On The Ecological Validity of a Password Study

Sascha Fahl, Marian Harbach, Yasemin Acar, Matthew Smith
Usable Security and Privacy Laboratory
Leibniz University Hannover, Germany
fahl, harbach, acar, smith@usecap.uni-hannover.de

ABSTRACT

The ecological validity of password studies is a complex topic and difficult to quantify. Most researchers who conduct password user studies try to address the issue in their study design. However, the methods researchers use to try to improve ecological validity vary and some methods even contradict each other. One reason for this is that the very nature of the problem of ecological validity of password studies is hard to study, due to the lack of ground truth. In this paper, we present a study on the ecological validity of password studies designed specifically to shed light on this issue. We were able to compare the behavior of 645 study participants with their real world password choices. We conducted both online and laboratory studies, under priming and non-priming conditions, to be able to evaluate the effects of these different forms of password studies. While our study is able to investigate only one specific password environment used by a limited population and thus cannot answer all questions about ecological validity, it does represent a first important step in judging the impact of ecological validity on password studies.

Categories and Subject Descriptors

Security And Privacy [**Human and Societal Aspects of Security and Privacy**]: Usability in Security and Privacy; Human-centered Computing [**Human Computer Interaction (HCI)**]: Empirical Studies in HCI

General Terms

Security, Human Factors, Measurement

Keywords

Usable Security, Passwords, Ecological Validity

1. INTRODUCTION

Passwords are the most common, widespread and possibly the most debated authentication mechanism in use. The inherent conflict of creating usable (e.g. user memorable) but secure passwords has kept security researchers busy ever since the introduction of passwords to computer systems in the 1960s. A lot of password policy and password advice is based on anecdotal evidence and theoretical security measures. However, particularly the last few years have seen an increasing number of academic studies into password security and usability. These studies can be divided into two major categories: studies of real world passwords (usually based on leaked/stolen password lists such as the RockYou and MySpace password databases) and user studies.

The obvious advantage of the first type of study is that the passwords in question are real and thus any results obtained from the study are based on accurate real-world data. However, these studies of course only shed light on the system the passwords were created in and do not allow researchers to experiment with different settings. As Kelley et al. [10] point out, there is also an ethical conundrum, since these password lists were obtained through criminal activity.

User studies offer the advantage of being directed by the researchers so different conditions can be used to study the effects of certain aspects of the password system, thus giving researchers the flexibility to study different security or usability aspects in a controlled situation. However, one great concern about user studies is the ecological validity of the study, i.e., do the study participants behave the way users would in real life and consequently, to what extent are the study results relevant and transferable to the real world?

In their recent work *Of Passwords and People*, Komanduri et al. [11] summarize this problem nicely:

> "It is difficult to demonstrate ecological validity in any password study where participants are aware they are creating a password for a study, rather than for an account they value and expect to access repeatedly over time. Ideally, password studies would be conducted by collecting data on real passwords created by real users of a deployed system. However, due to the sensitivity of password data and the difficulty of partitioning real users into experimental conditions [...] it is difficult to collect the data [...] from a deployed system."

To counter potential problems with ecological validity, researchers have developed different opinions on which form of user study offers the best ecological validity for a given research goal. Many researchers opt for online surveys to increase the sample size and diversity of their survey population. MTurk in particular has gained popularity.

> "Using MTurk allows us to study a larger volume of par-

Symposium on Usable Privacy and Security (SOUPS) 2013, July 24–26, 2013, Newcastle, UK.

ticipants in a controlled setting than would otherwise be possible" (cf. Kelley et al. [10]).

Buhrmeister et al. also state that the MTurk population is significantly more diverse than samples used in typical lab-based studies that heavily favor college-student participants [4]. Similarly, Bravo-Lillo et al. conducted a MTurk study for diversity reasons [3].

However, there are also online studies conducted using a more local population such as presented by Just et al. [9]. On the other side, Haque et al. chose a lab study over an online study because of results they obtained during a pretest [8]:

> "We conducted a laboratory experiment with 80 UTA students. Although a larger number of participants could have been drawn from an online study, we preferred a laboratory study because our pilot study (N=12) showed that a laboratory study would produce more consistent responses".

While there have been many user studies on a variety of aspects of password systems taking different measures to improve ecological validity, to the best of our knowledge there has been no study to examine the impact user study setups actually have on the ecological validity of these studies. In this paper we present a study evaluating several user studies in combination with real world data from our university. We conducted several user studies with students of our university who we asked to create passwords for services similar to their university services. With their consent, we then compared the study results to their real-world passwords for the same services using a number of different metrics.

Our results show that less than one third of our study participants created passwords that did not mirror their real-world behavior at all. Additionally, more than 25 % of participants actually used their real passwords during the study. We also find that the ecological validity of password studies can be improved by filtering participants using self-reported data and make recommendations for studies focusing on specific aspects of password usage.

2. BACKGROUND AND RELATED WORK

Ecological validity has been a concern for a great number of research projects. To the best of our knowledge, this is the first study concerning the ecological validity of password creation in user studies with the type of the study as the independent variable and with a within-subjects comparison with real world password data. However, ecological validity has been discussed in many password research papers. In the following, we present a brief literature overview of a selection of password and password system user studies with respect to the form of the study and the authors' thoughts on ecological validity. The list is far from complete, however, it gives the reader an overview of the vast spectrum of possible ways of running password user studies. We categorize the studies by the following attributes:

2.1 Description

It is believed that the description of a study can influence user behavior from the beginning. Some studies try to disguise their interest in passwords, hoping to not create a bias in their subjects: Haque et al. state:

> "We did not want to give the participants any clue about our experimental motive because we expected the participants to spontaneously construct new passwords, exactly in the same way as they do in real life" [8].

This sentiment is found in many studies. Another example is the work of Shay et al. [13]:

> "Ecological validity in many password studies is limited by the fact that participants are aware they are using passwords for a study, rather than for accounts they value or expect to use long-term".

A common approach is to ask participants to role-play a situation where password creation is just one step among others.

However, some studies openly state their interest in passwords or aspects of a password system [6, 7, 9, 14]:

> "Before beginning the experiment, participants were asked to pretend the passwords they create during the session were going to protect their online bank accounts, and they should create passwords that would be easy to remember but hard for other people to guess" [6].

Another paper added an interesting twist to this issue: Kelley et al. [10] had users set up one password for the study website, i. e., a real password, and subsequently had users role-play the creation of several further higher value passwords.

2.2 Study Type

Another aspect where different choices have been made is the type of study used. There are many options: The two most common choices are online and laboratory studies. However, there are also a few pen & paper based studies, as well as field and interview studies. Again, authors have different opinions on why they chose a particular type of study.

Many researchers opt for online studies [1, 10, 11, 13, 15]. Common reasons for this choice are the possibility of increasing the sample size and the diversity of the study population in comparison to laboratory studies conducted with students. However, there are cases where a paper-based survey was chosen instead of an online survey, for instance in the paper by Shay et al.:

> "While collecting and managing the data would have been easier online, we were concerned that more security-savvy users would be reluctant to provide truthful information if they thought we could link their responses to their usernames" [14].

They also reported: *"While pilot testing the survey, we received feedback that our password composition questions made respondents uncomfortable. Pilot testers expressed concern that we were gathering so much specific data about their passwords that we might be able to determine them. We feared that these concerns would prevent users from taking our survey or cause them to answer untruthfully"* [14].

The study by Just et al. used a combination of online and paper survey [9]. Just et al. intended to study the security and usability properties of the security questions that are commonly used when users forget their password. While the major part of the study was conducted online, the answers to the security questions were written on paper,

since the authors were worried that having the participants enter the security-critical answers online would prevent them from selecting realistic questions. They state:

> "Our experimental method presents an interesting option for obtaining more realistic authentication information in an ethical way. Though while the use of pen-and-paper aids us in this effort, the same practice introduces some factors that are difficult to control. For example, the self-assessment of memorability places a significant amount of trust in the participant" *[9].*

This problem is very closely related to studies requiring users to divulge realistic passwords.

A large number of studies use a laboratory setup to study password systems and password behavior (cf. [5, 6, 7, 8]). Laboratory studies have a number of well known issues that can potentially lead to ecological validity problems, such as the fact that users are not in their natural environment and are particularly aware that they are being studied. Some researchers have used this source of potential bias to try and err on the conservative side of their evaluation such as Haque et al. [8]:

> "Finally, we note that the presence of an observer may, if anything, motivate users to create stronger passwords than they might otherwise."

However, they could not capture and discuss whether or not this effect actually took place. Furthermore, if this effect does occur, it is not desirable for all types of studies.

Several researchers studying usability aspects of password systems avoid the problem of users potentially choosing unrealistic passwords by specifying the passwords themselves. This is often done in case the password itself is not the main focus of the study and the effect of password behaviors is limited for the sake of the study. Examples of this kind of study were done by Shay et al.[13] and Ur et al.[15]. However, even these studies must take ecological validity into account, since the usability of password systems is often effected by the strength of the passwords. For instance, Zakaria et al. preassigned passwords to the participants, but tried to match what users would choose:

> "In order to maintain ecological validity of this experiment, the passwords tested must be memorable; otherwise they would be less likely to be chosen in the real world" [18].

Researchers also attempt to analyze their data in a way that allows the detection of problems concerning ecological validity. Komanduri et al. state:

> "Two indicators that participants may have answered honestly are that their self- reported password reuse was higher in the basic survey condition than in the four other conditions, and that the computed entropy of passwords in these four conditions was significantly higher than the entropy of passwords in the basic survey condition. Both findings are consistent with users picking better passwords to protect a hypothetical email account than to protect a real survey account. Despite this, we cannot conclude that our results completely approximate real-world behavior; because the hypothetical scenario was the same across the four conditions, [...]".[11]

2.3 Ecological Validity

Schechter et al. [12] conducted a study on the ecological validity impact of personal risk and security priming in a phishing study. They conducted a between-subjects study with three groups. Two groups were asked to role-play a banking task. One of these was primed to pay attention to security while the other was not. The third group used their own personal data. Schechter et al. discovered that priming had no significant effect on the security behavior between the two role playing groups. However, there was a significant improvement in security behavior between the group using their personal data and the union of the role-playing groups. We found the same lack of effect of priming in our study and can offer additional insights into behavioral differences between using real and study data for the domain of password studies, as well as offering the new view of a within subjects design with ground truth data.

3. A STUDY OF STUDYING PASSWORDS

3.1 Preamble

In this paper, we present a study on the ecological validity of a password study to shed some light on this complex topic. We were in the fortunate situation of being asked for consultation by the Identity Management (IDM) team of our University's IT Services concerning their password policy system. In the course of this work, we discovered a unique opportunity: The IDM system stored up to five unique passwords per user using asymmetric cryptography, so it would be possible to decrypt the passwords to do a security analysis.[1] The passwords belonged to five university-wide services, comprising the identity management itself, eMail, Wifi, campus login, and Web-Single Sign On (SSO). Under the mandate to improve the security of our university's password system, we were provided with an anonymized dump of the decrypted passwords to help find policies that would prevent weak passwords without putting undue strain on the users.

However, in addition to this security analysis we were thus in the fortunate position to – in theory – be able to design a study that would mirror the enrollment process at our university and then be able to compare the passwords our study participants created to the passwords that they actually created for their real services. We therefore approached the Privacy Officer[2] with a suggestion for such a study. The study's goal would be to allow us to study the ecological validity of password studies based on this data. It would be prepared and run just like a regular password study in which we would ask the students to role-play the enrollment in an university's IDM system. As with all studies we would require informed consent from the participants at the beginning of the study to cover the study itself. However, the final question in the study would ask for an additional informed consent to allow us to compare the passwords our participants just provided with the passwords from their real accounts. Consenting to this comparison was optional and opt-in. We designed the study in such a way that we would never see the account information belonging to any

[1]While this is non-standard behaviour, this design choice was well-founded and is implemented securely.

[2]There is no formal IRB process at our university. The Privacy Officer also consults on ethical matters.

real or study passwords. The analysis of the real and study passwords would be conducted offline and without any demographic data. Only the results of the password behavior analysis would then be linked to the demographic information collected in the study. All results were to be checked with the Privacy Officer before publication. We discussed this study design and its legal and ethical ramifications in detail with the Privacy Officer and the IDM team. Since the study was based on informed consent and the comparison with the participants' real life passwords was covered by a second, separate and opt-in informed consent agreement, our study protocol was approved.

3.2 Study Design

Given the wealth of different questions about ecological validity we could try and answer, we had to pick a manageable number to fit into our study. The main question we wanted to answer was: Do passwords generated by participants asked to role-play a scenario in which they have to create a password for fictitious accounts resemble their real passwords? Or do participants behave so differently because of the study that the results of the study should not be used to make inferences about their real behavior? Based on our literature review the two prevalent forms of user studies for passwords are online and laboratory studies, so we decided to study these two forms of experiments.[3]

Since the password system of our university has password policies in place that force users to create fairly strong passwords,[4] we were concerned that the effect size might be fairly small since the policies rule out simple passwords. Thus we decided to add only one more independent variable to the mix and examine whether openly mentioning that the study is also about passwords has an effect compared to obfuscating the study's purpose. We selected this variable since many papers chose to invest a fair amount of effort to obfuscate their study, specifically stating a wish to avoid priming the subjects in the hopes of getting more realistic results. However, to the best of our knowledge, there is no evidence to suggest that this is a good approach. In fact, it may even be counterproductive.

Altogether, in addition to the two within-subjects conditions of real vs. study passwords, our study covered four between-subjects conditions in two variables (lab vs. online study; password priming vs. no password priming). In all conditions, we asked students to role-play that they had just enrolled in a new university and needed to register for the different services offered by the university. We used the same type of services as offered by our real university.

In both studies, we applied the same password creation policies that are currently enforced for IT service accounts at our university:

- A password's minimal length is 8 characters; its maximal length is 16 characters.
- Password characters are split into four different groups: Upper and lower case alphabetical characters, special characters `,.:;!?\#\%\$@+-/_><=()[]{}*` and digits. Passwords that are shorter than 12 characters must include characters from three of the four described character groups. Passwords that are 12 characters or longer must only include characters from two of the four described character groups.
- Neither the student's first/last name nor the student's ID number may be part of a password.
- Users must use different passwords for all accounts.

3.2.1 Online Study

We invited 16,500 university students via email to participate in our online study, announcing a two-part online study on the creation of online accounts for university services. Participants were told that each part consisted of a simulated online scenario combined with an online questionnaire, taking between 15 to 20 minutes and 5 to 10 minutes respectively. As incentive, we offered our participants the option to enter a raffle for three 100 Euro Amazon vouchers. The email also stated that the second part of the study would follow two days after the first and that they would be able to enter the raffle only after completion of part two.

To cover the second independent variable, we varied the introductory text. The invitation email was the same for all conditions, inviting students to participate in a study about online account enrollment at a university. After students clicked on the link to enter the study, two different introductory texts were shown. For the non-priming condition, the text just stated that participants should pretend they were enrolling in a new university and should behave as they would in real life. The word "password" was not used at all. For the priming condition, we mentioned that it was important to keep the passwords for the accounts available. We asked the participants to take exactly the same steps they normally take when creating and managing new passwords. We also asked the participants to act as if the passwords for the fictitious study scenario were real passwords. This is the same information about passwords that was used in Kelley et al.'s work [10].

In both conditions, participants were told to imagine that they just enrolled in a new university and intended to use different IT services. Therefore, accounts for the Identity Management System, Email, WiFi and the Campus login service had to be created. The description for both conditions stated that to complete the second part of the study two days later, it would be necessary to log into those accounts again. We included this condition since it is a common approach for researchers to try to urge participants to use passwords they would be able to remember/keep for a while, as opposed to single-use throw-away strings.

After setting up the accounts for the four services, participants for both conditions were redirected to an online survey. The online survey collected demographic information and information about the participants' Internet usage. We also asked the users how they usually manage their passwords. They were also asked how many different passwords they use for all their online accounts to self-report the quality of the passwords they created in the study they just completed compared to their real passwords and if their password creation behavior in the study was different from their behavior in everyday Internet usage.

After two days, our participants received a personalized email requesting their participation in the previously announced second part of the study. After clicking a link con-

[3] We did not use MTurk like many other studies have, since we would have lacked ground truth data to compare the behavior for those participants.

[4] As judged by the password policy system.

tained in the email, each participant was asked to log into the same four services as before, using the password they had created two days ago. After three tries, participants could choose to continue to the next service without successfully logging in, in order to not unnecessarily frustrate our subjects. The system recorded whether or not participants succeeded and how many tries each participant failed. Finally, participants completed a second questionnaire asking how they had managed the study passwords.

3.2.2 Lab Study

We also invited 740 university students to a lab study from our study mailing list. We excluded them from the invitation to the online study, so they did not receive two invitations to this study. Our goal was to conduct a lab study with roughly 70 participants so we invited 740 students, since we usually have a response rate of 10%. We arranged appointments with 75 students of which 68 actually attended the lab study. The study was set up the same way as the online study, the only difference being that the students had to complete the password creation for the study in an unfamiliar lab environment, with a lab computer and under the supervision of the experimenter. After a brief welcome speech, the lead experimenter read an introductory description of the study aloud equivalent to the online study's description.

While the first part of the study was conducted in our usability lab, the second part could be completed two days later at home. Again, we sent out personalized invitation links for each participant. The participants were told that they would receive 20 Euros each after they completed both parts of the study. Before the first part started, participants had the chance to ask questions or make comments. They were also told that they could request assistance if they had technical difficulties with the lab computer.

4. PASSWORD ANALYSIS

Analyzing passwords is a hotly debated topic. Since our main interest was not in a particular measure of password strength but in researching user behavior, we decided to begin with a manual scoring of different password metrics/patterns.

4.1 Expert Scoring

The goal of our manual scoring was to categorize participants based on how similar the metrics of their study passwords were compared to their real passwords. We decided to use this type of review instead of a more algorithmic approach to be able to accommodate the nuanced differences in user behavior that are difficult to capture using formalized rules. Additionally, to the best of the authors' knowledge, there has not been any work on directly comparing user passwords to judge their behavior. We therefore favored a manual approach to explore this aspect. For instance, using metrics alone, it would have been difficult to catch the different behavior for the following fictitious example passwords: Study: "PwdIDM11.", "PwdMail11.", "PwdWifi11.", "PwdPC11." and Real: 'B0ru$$ia09", "16.Januar", "(australien)", "314159Pi". As can easily be seen, the study passwords follow a clear system while the real passwords don't. However, the bit-strength of the password (as calculated according to an approximation of Shannon[5] and NIST[6], respectively) and the crackability (as calculated according to John) is fairly similar. While it would of course be possible to create custom metrics to try and factor in similarity between the password groups, the options would have been endless and unverified. We would also not have been able to capture behavioral anomalies encoded in the passwords such as these real examples from one participant in the study: "studiesSuck123" and "IamSoBored!!!". Since the participant's real passwords did not use such references, this is a case where the attitude in the study differs from the behavior in real life.

To capture password behavior, we define the following metrics and guidelines aiming to capture a general idea of user behavior instead of pure password strength in an expert scoring process. We break down a password into the following components:

Names Any kind of name, i.e. persons, nicknames, pets, places, etc.

Dictionary Word Any word contained in a dictionary.

Dates Any kinds of date, no matter the form or length, e.g. `1999, 02/03/13, 78, Feb.2., 09081978`.[7]

Simple Numbers Single numbers, counters such as `1,2,3,4` or simple sequences such as `123,456` or `111,222` etc.

Complex Numbers Any combination of numbers that are not dates or simple numbers.

Lower Case String String containing only lower case characters.

Upper Case String String containing only upper case characters.

Mixed Case String String containing mixed case characters.

Special Characters Any combination of special characters.

L33T Speak The use of leet speak.

Keyboard Pattern A combination of characters arising from using adjoining keys on common keyboards, e.g. `QWERTY, sdcx, 7895123` etc.

Random String A random string containing letters, numbers and special characters that could not be sensibly broken down into the categories above, e.g. a string generated by a password generator. We might have misinterpreted strings as random although they followed a structure such as the first letters of words in a sentence.

We considered the following transformation rules to judge the similarity of one password to another:

[5]cf. Mathematics of Information and Coding, Chapter 2

[6]cf. `http://csrc.nist.gov/publications/nistpubs/800-63/SP800-63V1_0_2.pdf`

[7]In some cases it was hard to differentiate between numbers and the year in the yy form. In these cases we used the context of other passwords of that participant to try and score correctly.

Ordering and Reordering of Components The order in which components are used

Exchange of Content Any exchange of an instance of a component by different instance of the same component.

Incrementation Any form of systematic incrementation, e.g. `3,4,5` or `!, !!, !!!`

Changing of Case Any changes between upper and lower case for a component or parts of components.

Insertion An insertion of one component into another, e.g. `password` and `1234` to `pa1ss2wo3rd5` or `password` and `...` to `pa.ss.wo.rd`.

Using these guidelines, each of the authors scored all participants that had given us permission to compare their study passwords with their real passwords. To assist the manual scoring, we preprocessed the passwords and appended a " [kb]" to passwords which contained a keyboard pattern.

Each subject's password set was assigned to one of the following categories:

Null No apparent similarity between the real passwords and the study passwords

Single There is one study password that is similar to a real password, however, the sets are not similar to each other

Full There are several study passwords that are very similar to real passwords and there is a similarity between the sets as well

System There is a system within each set and the systems are similar, but the composition of the passwords between the sets is not the same

Derogatory Obvious and derogatory reference to the study indicating that the participant did not show normal behavior

The difference between categories Full and System is fairly small. One additional criterion for category System is: If shown eight passwords in random order, is it possible to distinguish two sets of four passwords? If all eight passwords are so similar that it is impossible to distinguish between the sets, the subject is scored as category Full. This scoring system is not designed to measure password strength but likeness/provenance. To put it differently: How useful and accurate are the passwords given in the experiment to study the real life behavior of our participants? This scoring (on its own) does not take strength into account, i.e., rose123 would match Elisabeth9876. We combine this metric with password strength at a later stage.

Table 1 shows some examples of our scoring system. The passwords shown there are inspired by real cases, but have been altered so as to not endanger any real user accounts. We discuss each example in the following:

1. In this example, the passwords explicitly reference the study. Since the real passwords are not similar we score this as Derogatory.

2. The real passwords contain a single letter plus a date and a long upper case string with some numbers. Three of the study passwords contain several dictionary words. There is no similarity of pattern at all, so we score this as Null.

3. Two of the real passwords are dates and two are manglings of the same name. The study passwords are dictionary words with numbers and some special characters thrown in, so this is also scored as Null.

4. Both sets of passwords are based on a name plus a number. However the real set is based on name plus date and the name is varied between the passwords. In the study passwords, the name is the same and instead of a date a simple number is used that is incremented over the password. The study passwords are more homogeneous than the real passwords. Thus a singular password is similar but the overall behavior between study and real life is visible and thus this set is scored as Single.

5. This set is the reverse of the above. The study passwords are more heterogenous but there are singular passwords similar to singular passwords in the real set.

6. In these sets all passwords are very similar. They all use a base word and a sequence of numbers. Thus the set is scored as Full.

7. The same goes for this set. All passwords are of the same nature, i.e. a random string, so this the set is scored as Full.

8. These two sets are generated using the same principle. Both are based on a word plus a number. There are slight differences though. The real passwords are based around variations of the number 99499 while the study passwords use a number pattern with an increment. Thus while no two passwords are the same, it is plausible that the same user created them and it would easily be possible to sort the four correct passwords into the two sets. Thus this set is scored as System.

9. The system in these passwords is also clearly visible. The participant uses the numbers 3 and 5 and a base word and alters spelling and order between passwords. While the similarity is not as high as in the last example it still seems plausible behavior for the user and thus is scored as System.

4.1.1 Scoring Conflicts

Each of the authors scored the entire dataset separately and in a different order. The scores of all three raters agreed entirely in 47.2 % and disagreed entirely in 9.3 % of cases. The remaining 43.5 % of conflicts had two scores agreeing and could have been solved using majority votes. However, we decided to discuss each participant we did not fully agree on individually. These discussions were conflict free and were usually resolved by the majority explaining the pattern or lack thereof. The final count of the categories is presented in Section 5.4.

Table 1: Examples of the expert scoring process for passwords. The first line in each row are passwords provided in the study while the second shows corresponding real passwords.

#	IDM	Mail	Wifi	Campus PC	Score
1	notsecure12 TreePeter$1	ihatesurveys77 woJIJIui	2moreforyou TreePeter$2	Iknowwhatthisis4	Derogatory
2	EifkLegs KDOSKDO2EWKFD2	CornFlakes	YeaYeaYo U03.03.12	Mineralwater	Null
3	Saver3451 9thFeb90	Lions.Den54 Feb9th90	Plants1.go Peteeer1	Soon,me.1 Peteeeer2	Null
4	Roses220 Mary0908	Roses221	Roses222 Physics2010	Roses223 Maths2010	Single
5	Intovgaad! Intovgaad!	Sydney12 Intovgaad!!	Spain13 !Intovgaad	Hello123 Intovgaad?	Single
6	Fryingpan123 Fryingpan123	Fryingpan456	Fryingpan789 Fryingpan456	Fryingpan99999 Fryingpan789	Full
7	9;6BU7MG3h#y #M24kJB	d<8k@L343oju 38333DI(*DL33T	s$jW7Q639C)H B[L72:7L7cvA	KcL4.,8b7T4A	Full
8	Unlockthis1122655 Secret99499	Unlockthis2233766	Unlockthis3344877 Secret994	Unlockthis4455988 Secret99499!	System
9	Jumpman35 3kefdUed	5JumpmanThree Three5fun	FiveJumpman3 three5Fun	5Jumpman	System

4.1.2 *Interpretation of Scores*

The usefulness of participants for a password study will depend on the research focus of the actual study. If password behavior needs to be studied over several services or passwords, participants in categories Full and System are useful. Our feeling was that participants in categories Full and System both behaved realistically with participants from category Full having more similar passwords in general. While category Single participants can still add value, they can also introduce unrealistic behavior: For instance, they show heterogeneous behavior in the study but have homogeneous passwords in real life or vice versa.

If only a single password is to be studied, our feeling is that participants from category Single are probably acceptable to study. However, it should be noted that the matching password was not always the first participants entered. There were cases where it seemed that the participants used up throw-away passwords until they ran out and then used a real password. However we could not measure this in any meaningful way and thus this feeling should be taken with due caution. Participants in categories Null and Derogatory did not behave consistently and could skew the results of a study in a damaging way.

Apart from our manual scoring, we also applied some traditional password metrics for further analysis and to support our scoring.

4.2 Password Composition

Above, we analyzed the structure of the passwords manually and with a fairly coarse granularity. Another measure of similarity for passwords is their composition with respect to single characters. Therefore, we calculated the following composition metrics for every password from both the real accounts and the online and lab study accounts:

- The length of a password.
- The number of upper case characters.
- The number of lower case character.
- The number of digits in the password.
- The number of special characters as defined in the deployed password policies (cf. Section 3.2).
- An approximation of the Shannon entropy for the password.
- The NIST entropy for the password.

In addition to the above metrics we also analyzed our password corpus for the same patterns as described in Section 4.1 algorithmically. For the dictionary check, we compiled a dictionary based on multiple wordlists. These wordlists include Burnett's top 10,000 passwords,[8] lists of first and surnames taken from Wiktionary,[9] an English and a German dictionary, the top 10,000 German words,[10] a list of 85 common emoticons and the following list of study specific words we compiled based on service names and other prominent words. Our algorithm then checked if a password or parts of a password could be matched against the dictionary. Additionally, our algorithm analyzed passwords for the occurance of leet speak. Leet characters were translated into non-leet speak, then we checked if the translated version could be found in the dictionary. Example: `W@llcOlO2` is first translated to `Wallcolor` and then both `wall` and `color` could successfully be matched against the dictionary.

4.3 Password Strength

Password strength has been measured in many different ways: From simple 0 entropy, to more elaborate bit strength metrics, guessability and resistance against cracking attacks [2, 10]. There is a fair amount of discussion going on about which metric gives the most realistic measure of password strength for a given type of attacker. In this study, the password strength aspect plays a secondary role since we

[8] cf. http://xato.net/passwords/more–top–worst–passwords
[9] cf. `http://en.wiktionary.org/wiki/Appendix:Names`
[10] cf. `http://wortschatz.uni-leipzig.de/Papers/top10000de.txt`

are mainly interested in the relative comparison between the sets of passwords generated by the same user. We therefore chose the following measures:

4.3.1 Entropy

To compare the relative strength of participants' real and study passwords, we chose two well-known entropy measures. We used an approximation of plain Shannon entropy, i. e., $H = log_2 N^L$ where N is the number of symbols in the alphabet the password is based on and L is the password's length. This approximation of plain Shannon entropy has been repeatedly criticized [2, 10] to not accurately represent a password's strength against an attacker. However, in our case, we were interested in comparing the relative information content of several passwords created by the same user. To this end, the approximation of Shannon's entropy represents an upper bound of the potential information content of passwords. Furthermore, we also applied the NIST entropy [16] for passwords to get a more conservative estimate of a password's information content. The NIST entropy estimate limits the influence of password length and the use of different character classes while providing an easy to compute set of rules. In both cases, we do not suggest that these measures represent a good measure of absolute strength of a password. We merely wish to compare the values between the study and real datasets.

4.3.2 Crackability

We also compared password strength by subjecting each set of passwords to dictionary attacks using the well-known password cracker "John The Ripper". For all sets, we used three dictionaries: the dic-0947 dictionary that has shown good password cracking performance in related work [16, 17], a list of 220,000 German words from LibreOffice's spell checker, and the over 14 million stolen passwords from the RockYou set which has also been often used [2, 10, 15, 16]. In a second run, we also used the study passwords as a wordlist against the participants' real passwords. Each wordlist was additionally mangled using 1,080 rules from John's "*Single*" ruleset [16]. For the subsequent analyses, we compared how many passwords per subject were crackable using these attacks.

5. RESULTS

5.1 Participants

Overall, 765 participants participated in our online study and 68 in our lab study. The first 500 respondents in the online study and the first 35 in the lab study were assigned to the priming conditions. Altogether, 75.7 % (579) of all online participants and 95.6 % (65) of all lab participants completed part two of the study.

We removed the following participants from our evaluation: 85 online and 3 lab participants who did not give their consent that we may compare their real passwords with their study passwords, 8 online and 2 lab participants who did not supply a valid student ID and thus we could not obtain their real passwords and 53 online and 1 lab participant(s) that had only one real password with the IT services department to base our scores on. Since some participants matched criteria in multiple exclusion categories, this left us with a total of 645 records (583 online and 63 lab). Of the 583 participants in the online study, 66% were exposed to the priming condition and of the 63 participants of the lab study 53% were exposed to the priming condition. Across all conditions, participants were aged between 17 and 55 (23.72 years on average, $sd = 4.31$, median=23), 35.8 % were female, 16.3 % studied an IT-related subject. Participants self-reported medium IT expertise (average score 3.42, $sd = 1.0$, median=3 on a five point scale anchored at 1=high IT expertise and 5=low IT expertise). The majority of respondents stated that they use the Internet repeatedly throughout the day (90.7 %). They reported an average of 18.1 online accounts ($sd = 21.0$, median=14). 17.4 % had account credentials abused at least once before, only 42 (6.5 %) had never forgotten a password before. The majority (79.6 %) had forgotten a password at least twice. 63.2 % respondents used between 2 and 5 passwords for most of their online accounts and 14.9 % used different passwords for all accounts. Participants' passwords in the university IT services database had an average age of 534 days ($sd = 391.7$ days, median=481). 26.5 % used at least one of their real passwords in our study.

Due to a technical problem in condition assignment, participants were not assigned to conditions in a round robin process but sequentially. This had two undesirable effects: first, the non-priming condition in the online study had fewer participants than the priming condition and, second, the average age of the real passwords is lower in the lab study than in the online study (551.2 days online vs. 370.4 days for the lab, medians: 502 online vs 246 for the lab; Kolmogorov-Smirnov-Test for equality, one-tail, alternative=less: $p = 0.0001817$). We tested if removing older passwords would have an effect on any of our tests, but did not find a significant difference. We did not find any demographical differences across our four between-subjects conditions. While the smaller N for the prime-online condition may diminish the sensitivity of our statistical tests, the overall number of participants in the online conditions is large enough to compensate for this. We did not find any significant effects of password age on the password metrics introduced above and could not find any other indication that this confound effected our results.

5.2 Scoring Evaluation

The first step in our evaluation was to check whether our categorization described the relationship between the real and the study passwords correctly. We had several hypotheses concerning the correlations we would find in the different categories: category Full participants would have the highest correlation of password composition values between their two password sets of all categories. We expected a weaker correlation for category Single and category System participants and no correlation for category Null and Derogatory participants.

To evaluate our scorings and the hypotheses above, we conducted Kendall's Rank Correlation Tests for all password composition values presented in Section 4.2, the entropy measures introduced in the previous section between the study and real password set as well as the crackability of the passwords. As expected we found highly significant and strong correlations for participants in score category Full and mostly significant correlations in categories Single and Systemas can be seen in Table 2. However, it needs to be noted that while we found significant correlations for those three categories, we found no correlation when the entire set

Table 2: Password Metrics Real vs. Study (Kendall's τ).

	Derogatory		Null		Single		Full		System	
	τ	p	τ	p	τ	p	τ	p	τ	p
Length	.5352	.0464	−.0439	.3994	.2157	.0008	.5141	< .0005	.0581	.6492
Shannon approx.	.6111	.0247	−.0368	.4609	.2006	.0012	.4768	< .0005	.0038	.9753
NIST	.1538	.5854	.0778	.1311	.0022	.9731	.2884	< .0005	−.1413	.2564
Digits	−.1492	.5923	.0762	.1523	.3686	< .0005	.6528	< .0005	.2577	.0541
Upper Chars	.4620	.1030	.1830	.0009	.2584	< .0005	.5779	< .0005	.1451	.2908
Lower Chars	.1714	.5272	−.0133	.7954	.3100	< .0005	.6095	< .0005	.1200	.3413
Special Chars	.6365	.0301	.3853	.0005	.5376	< .0005	.6482	< .0005	.3733	.0095
Crackability	.7324	.0250	.1066	.1126	.3352	< .0005	.5514	< .0005	.0755	.6465

We conducted a correlation test within the categories, comparing study password sets with the respective real password sets. We applied the Bonferroni correction that gave us an alpha value of 0.0063. As expected, we found highly significant correlations in category Fullsome significant correlations in categories Single and System and rather random correlation behavior in categories Derogatory and Null. This strongly supports our scoring procedure, while also pointing to the limits of assuming the correlation of the above metrics to be very strong between studies and real passwords.

of study passwords was analyzed as a whole.

We found no correlation for the categories Null and Derogatory.

To simplify the further evaluation, we conducted tests to see whether we can legitimately speak of Single, Full and System participants, regardless of the condition (online or lab, priming or non-priming) they were in. For this we conducted 2-tailed Kolmogorov-Smirnov tests which are documented in Tables 6 and 7 in the Appendix. The results show that there was no difference between those conditions with respect to our categorization and thus it is possible to compare the differences in password behavior solely on the category irrespective of the condition.

This shows that our scoring was consistent: Participants classified to behave consistently between real and study passwords by our scoring system did compose their passwords consistently, while those deemed to behave inconsistently according to our classification indeed produced independent sets of passwords.

This leads us to assume that category Single, Full and System participants behave more realistically in our study than category Null and Derogatory participants, with category Full participants showing the strongest correlation. 26.5% of our participants even used at least one of their real passwords in the study. In the following we refer to the combination of categories Single, Full and System as helpful passwords and the combination of categories Derogatory and Null as unhelpful passwords - in the sense of helpful or not helpful to study realistic user behaviour.

5.3 Evaluation

Across all conditions, we found that we had scored most password sets - 46.2 % (298) - into category Full i.e., as being very useful for studying password behavior. We assigned 18.8 % (121) password sets to categories Single and 5.1 % (33) to category System respectively, both in our opinion still representing partially valuable password samples. 28.5 % (184) password sets were assigned to categories Null and Derogatory (1.4 %), respectively, i.e., passwords that showed abnormal and derogatory behavior. In the following, we will compare how the different conditions affect the results based on this categorization.

5.4 Online vs Lab Study

Separating our scoring results by the type of study reveals a trend towards more realistic results in our lab study: More participants fell into the helpful categories Single, Full and System compared to our online study (cf. Table 3), the trend being significant according to Fisher's Exact Test ($p = 0.0296$ cf. Table 14). These results add weight to Haque et al.'s 12 participants pilot-study's observation that a laboratory study would produce more consistent responses [8]. While these results are statistically significant for our study, this should not be generalized without care. Please check the limitations discussed in Section 6 for more information on this.

5.5 Priming

Separating our scoring results by the priming and non-priming condition did not show a meaningful difference (c.f. Table 3). We verified this by performing Fisher's Exact Test on the 122 primed vs 71 non-primed unhelpful password sets and the 300 primed vs 152 non-primed helpful passwords sets. The null hypothesis that there was no difference in behavior could not be rejected with $p = 0.4698$ (alternative=two-tailed).

5.6 Self-Reported Values

We went on to evaluate which self-reported metrics of participants may serve to predict inconsistent study behavior. First of all, we directly asked participants if they behaved differently during the study. Participants that reported different behavior showed significantly fewer counts in categories Full, Single and System and higher counts in category Null and Derogatory as seen in Table 10. Whether or not a participant failed to remember their password after two days did not have a significant impact on the scores distribution as seen in Table 13 and neither did participating in the second part of the study as seen in Table 11.

Overall, participants who changed their usual behavior for the study obtained significantly fewer ratings in categories Full, System and Single and more in Null and Derogatory than participants who did not self-report this, as can be seen in Table 10. Finally, participants who said that they use individual passwords for each account also scored significantly more frequently in categories Null and Derogatory when participating online (cf. Table 12).

We also manually analyzed the reasons participants gave

Table 3: Scoring Results Online vs. Lab, Priming vs. Non-Priming

Score	Total		Online		Lab		Priming		Non-Priming	
Derogatory	9	(1.4 %)	9	(1.5 %)	0	(0 %)	4	(0.9 %)	5	(2.5 %)
Null	184	(28.5 %)	172	(29.5 %)	12	(17.9 %)	118	(28.0 %)	66	(29.4 %)
Single	121	(18.8 %)	108	(18.7 %)	13	(20.6 %)	80	(19.0 %)	41	(18.4 %)
Full	298	(46.2 %)	267	(45.8 %)	31	(49.2 %)	199	(47.1 %)	99	(44.3 %)
System	33	(5.1 %)	26	(4.5 %)	7	(11.1 %)	21	(5.0 %)	12	(5.4 %)

for deviating from their normal behavior. We found the following categories:

Disclosure Participants stated that they did not trust us or did not trust others with their real passwords in general.

Memorability Participants stated that they chose simpler passwords because otherwise they would have problems remembering them.

Value Participants stated that they chose simpler passwords because the passwords were unimportant. There was often a reference to it being "only a study".

Overburdened Participants stated they were overburdened by having to choose four passwords in short succession.

Policy Participants stated that they chose stronger passwords than normal because the password policy forced them to.

Lazy Participants stated that they were too lazy to choose proper passwords, or that they just wanted to get through the survey as quickly as possible.

New Behavior Participants stated that they adopted a new way of creating passwords in general and thus their old passwords were different.

None of the specific reasons for changing password behavior listed above had a significant influence on the participants' categorization as compared to the total of participants who admitted to having changed their behavior for the study.

5.7 Consenters vs. Non-Consenters

Altogether, 88.6 % of all online participants and 95.6 % of all lab participants gave their consent to compare their real passwords with the study passwords. We analyzed if participants who did not consent to the comparison with their real passwords showed any demographic deviations from the ones who did consent. We only found that those participants reported to have different passwords strategies: They stated that they use individual passwords per account more frequently, as shown in Table 12. We performed two-tailed Kolmogorov-Smirnov tests to see if study passwords supplied by participants who consented to our comparison with their real passwords have similar metrics as the study passwords of non-consenters. The above p-values suggest that there are no statistically significant differences between the two samples for the measured metrics.

Table 4: (Study) Password metrics for Consenters vs. Non-Consenters (2-tailed Kolmogorov-Smirnov).

	P-Value
Length	$p = 0.6183$
Shannon approx.	$p = 0.5852$
NIST	$p = 0.9408$
Digits	$p = 0.6352$
Upper Chars	$p = 0.0648$
Lower Chars	$p = 0.3119$
Special Chars	$p = 0.9803$
Crackability	$p = 0.9895$

5.8 Participants vs. Non-Participants

Due to the nature of our password ground truth data, we can also estimate how well our study participants represent the entire population of students to a certain extent. Since our university's IT services provided us with an anonymized set of passwords for all students enrolled for IT service accounts. We calculated average password length, entropy measures, the number of upper, lower and special chars and digits for this set and the set of students that participated in our study. We then conducted 2-tailed Kolmogorov-Smirnov tests for all metrics (cf. Table 5).

Table 5: (Real) Password Metrics for Participants vs. Non-Participants (2-tailed Kolmogorov-Smirnov).

	P-Value
Length	$p = 0.1329$
Shannon approx.	$p = 0.5005$
NIST	$p = 0.7400$
Digits	$p = 0.1623$
Upper Chars	$p = 0.7928$
Lower Chars	$p = 0.3494$
Special Chars	$p = 0.6344$
Crackability	$p = 0.4181$

These results suggest that there is no statistically significant difference between both participants and non-participants and hence we believe that our study sample adequately represents our university's population. Summaries of entropy and crackability for both participants and non-participants can be found in Table 8 and 9.

6. LIMITATIONS

Our study is limited in several ways.

Population: Since our ground truth data was drawn

from the student population of our university, our study also focused solely on this population. While this offers a certain amount of transferability to similar studies, the results should be used with care when evaluating the behavior of a more diverse population.

Password policies: Due to the policies in place a certain minimum password quality was enforced. Thus, the range across which participants could behave differently was restricted. Hence, it is possible that different behavior would be more pronounced in unconstrained password creation scenarios. However, since many password systems have policies in place, we believe this to be only a minor limitation in practice.

Self-selection bias: All our participants were self-selected. While this would constitute an ecological validity problem if these results were to be transferred to the general population, we believe in this study it is not a problem, since the matter we are studying (i. e. password studies) usually have the same self-selection procedure and thus our results should be accurate in this respect. Additionally, we were able to show that in our case the measured metrics of the passwords of participants and non-participants did not differ significantly (c.f. Table 5).

Number of real services: Not all students were registered for all real services. Consequently we might have missed behavioral patterns that would have become visible if we had been able to analyze more of their passwords. Potentially this could have upgraded a category Single participant to a Full or System.

Study enrollment vs real enrollment: We expected our participants to enroll in all four services in short succession. While this is not unrealistic per se, the enrollment process at our university does allow students to add services at a later date. There were no logs available to indicate how many students enrolled for all their services when they first signed up and how many added services over time. If students changed the way they choose their passwords between the enrollment for different services, we might have falsely classified a real category Full or System participant as a Single. Four participants also stated that they had felt overburdened by having to choose four passwords in a row.

Changing behavior over time: The quality of our study could be negatively influenced by a varying amount of time between the last time a participant changed their real password and participation in the study. If a participant genuinely changed the way they create passwords, e. g. adopted the use of a password manager or opted for a different method of designing multiple passwords, we might have misclassified a category Single, Full or System participant as a Null. However, we did not find any significant differences in our ratings based on the age of the user's real passwords. Five participants stated that the reason their study password differed from the real university passwords was due to the fact that they had changed the way the create password in general.

Different Incentives: We offered online study participants to enter a raffle for three 100 Euro Amazon vouchers, while each lab study participant received 20 Euros immediately. This might have influenced their motivation to put effort into thinking up sensible passwords, which might have contributed to differences in our findings between the two groups. However, since this mirrors our behavior when conducting real studies, this is an effect we would also encounter in future real studies.

Priming Due to a technical problem in condition assignment, participants of the online study were not assigned to the priming/non-priming condition in a round robin process but sequentially. We checked for both demographical and study result differences (as discussed in section 5.1) but we did not find any indication that this issue affected our results. A further possible confound is that students assigned to different conditions might have communicated about the study before participating and thus affected the non-priming condition.

Overall, although our dataset is not ideal, we contend that our findings do provide significant insight into the ecological validity of password studies. Since very little is known about this important topic, even imperfect information offers valuable insights at this stage.

7. CONCLUSIONS AND FUTURE WORK

In this study we presented an empirical analysis on the ecological validity of a password study. We manually compared 645 sets of passwords collected in an online and a laboratory study with real passwords belonging to the same participants for the same kind of services. We classified participants into five categories depending on how closely their study behavior matched their real behavior. We showed that our classification was a good predictor of positive correlation between a number of other password composition metrics as well as a password cracking count produced by John the Ripper. Based on these metrics, we estimate that 29.9 % of our participants did not behave as they normally do, while 46.1 % percent offered comparable data and 24.0 % offered somewhat comparable data. This improves to 19.6 %, 57.3 % and 23.1 % respectively after removing the participants who self-reported that they did not behave normally. To the best of our knowledge, these are the first empirical results on how people's password behavior changes due to the fact that they are participating in a password study.

7.1 Take-Aways

- A noteworthy number of study participants (26.5%) used one of their real passwords in the study. Beyond these direct matches, there were many study passwords that were very similar to participants' real passwords. Consequently, passwords gathered during a study should be treated with the same level of protection as real passwords. Normally, we analyze data collected during our studies on our laptops. For this study, we opted to work in encrypted volumes on computers disconnected from the network and all study related data has now been put in an encrypted drive which is stored in a university safe. We will adopt this procedure for all future password studies, due to the considerable number of participants who used their real passwords during the study.

- While there are participants who do not behave realistically during password studies on the whole, we argue that password studies create useful data to study. However, since real password studies do not know which participants are behaving normally and which are not, more research is needed to find out how to best interpret the results. Great care should be taken when

comparing a whole set of study passwords using standard metrics such as password length or NIST since the results can be noticeably skewed by the unrealistic behavior of the Null and Derogatory participants.

- More participants fell into the helpful Single, Full and System categories in our lab condition compared to our online condition. This difference is statistically significant.
- The difference between the priming conditions was minimal. There was no significant difference in our scoring. The slight differences in the NIST entropy were not conclusive.
- In our study, there was a relation between those participants we ranked as Null or Deregatory and those who self-reported they did not behave realistically. While this phenomenon needs to be studied in more detail and with different populations, it seems that adding this kind of self-reporting question to password studies can improve the quality of the data to a certain extent.
- Studies wishing to examine the memorability of passwords need to pay the most attention to ecological validity, since we saw a significant variation between users' normal behavior and their study behavior in respect to writing down passwords and selecting passwords to be memorable only for the duration of the study. Using online studies, participants are able to use all their normal means, i. e. writing passwords down, password managers etc. Conversely, however, a significant number of participants wrote down passwords although they stated they normally don't. The lab condition on the other hand hindered participants who normally wrote down their passwords from doing so. The lab condition also had a significantly higher login failure rate for part two of the study. If brain powered memorability is to be studied, we would recommend a laboratory study over an online study.

This study represents a first step to understanding the effect ecological validity issues have on password studies. There are several important and interesting open questions. One of the most relevant questions for future work is whether MTurkers behave in a similar way to the student population studied in this work. Since we have no ground truth data for MTurkers, other methods for establishing this will have to be found. Another interesting question is how participants behave when not constrained by password policies. While many password systems do use policies, it would nonetheless be interesting to know if there is an additional risk to the ecological validity of studies that do not use password policy enforcement. Our evaluation of the self-reporting data suggests this is likely to be true. Further research on how to optimize the removal of unsuitable participants using self-reported data is also an interesting avenue to follow.

8. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[2] J. Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Proc. IEEE S&P*, 2012.

[3] C. Bravo-Lillo, L. Cranor, J. Downs, S. Komanduri, S. Schechter, and M. Sleeper. Operating System Framed in Case of Mistaken Identity: Measuring The Success of Web-based Spoofing Attacks on OS Password-entry Dialogs. In *Proc. ACM CCS*, 2012.

[4] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, Feb. 2011.

[5] S. Chiasson, R. Biddle, and P. C. Van Oorschot. A second look at the usability of click-based graphical passwords. In *Proc. SOUPS*. ACM, July 2007.

[6] A. Forget, S. Chiasson, P. C. Van Oorschot, and R. Biddle. Improving text passwords through persuasion. In *Proc. SOUPS*. ACM, July 2008.

[7] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proc. SOUPS*. ACM, 2006.

[8] S. M. T. Haque, M. Wright, and S. Scielzo. A study of user password strategy for multiple accounts. In *Proc. CODASPY*. ACM, 2013.

[9] M. Just and D. Aspinall. Personal choice and challenge questions: a security and usability assessment. *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 8, 2009.

[10] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proc. IEEE S&P*, pages 523–537, 2012.

[11] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *Proc. CHI*. ACM, 2011.

[12] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *Proceedings of the 2007 IEEE Symposium on Security and Privacy*, SP '07, pages 51–65, Washington, DC, USA, 2007. IEEE Computer Society.

[13] R. Shay, P. G. Kelley, S. Komanduri, M. L. Mazurek, B. Ur, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proc. SOUPS*, page 7, 2012.

[14] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proc. SOUPS*, volume 10, 2010.

[15] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. Mazurek, T. Passaro, R. Shay, T. Vidas, and L. Bauer. How does your password measure up? The effect of strength meters on password creation. In *Proc. USENIX*, 2012.

[16] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. ACM CCS*, pages 162–175, 2010.

[17] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek. Password Cracking Using Probabilistic Context-Free Grammars. In *Proc. IEEE S&P*, pages 391–405, 2009.

[18] N. H. Zakaria, D. Griffiths, S. Brostoff, and J. Yan. Shoulder surfing defence for recall-based graphical passwords. In *Proc. SOUPS*, page 6, 2011.

APPENDIX

A. SELF REPORTING

We asked our participants to self-report on several aspects of their password usage behavior using the following questions (translated from German):

Which usage behavior concerning passwords for Internet services best mirrors your behavior?
Please select one of the following answers: I use exactly one password for all of my accounts.; I use between 2 and 5 different passwords for all my accounts.; I use between 6 and 10 different passwords for all my accounts.; Each of my accounts has a unique password.; Other

Please specify how you keep track of your passwords.
Please select all appropriate answers. I memorize all my passwords.; I came up with a scheme that allows me to deduce the password for the respective service whenever needed.; I wrote my passwords onto a piece of paper stored in a safe place that I consult if needed.; I am using a password manager that stores my usernames and passwords for me.

Please select the appropriate answer for each statement.
Rate your agreement from "I agree completely" (1) to "I disagree completely" with the following statements: The passwords I created are similar to my real passwords.; I chose a completely different type of password than I normally would.; The passwords I created are less secure than my real passwords.; The passwords I created are more secure than my real passwords.

Table 6: Priming vs. Non-Priming (2-tailed Kolmogorov-Smirnov; P-Values).

	Derogatory		**Null**		**Single**		**Full**		**System**	
	real	study	real	study	real	study	real	study	real	study
Length	0.2857	0.9883	0.6353	0.9145	0.1373	0.0854	0.4663	0.4859	0.6160	0.9132
Shannon approx.	0.7460	0.9683	0.2639	0.4083	0.1072	0.1521	0.5290	0.7270	0.9445	0.7264
NIST	0.1641	0.9483	0.8775	0.9934	0.0117	0.5886	0.8292	0.6100	0.9445	0.9445
Digits	0.9483	0.9483	0.4571	0.7415	0.1394	0.9529	0.7030	0.4342	0.2177	1.0000
Upper Chars	0.4005	0.9883	0.7442	0.9993	0.7683	0.4521	0.9996	0.7453	0.9680	0.9445
Lower Chars	0.5121	1.0000	0.9163	0.9091	0.3403	0.1865	0.1329	0.6774	0.6714	0.9838
Special Chars	0.5121	0.9991	0.1412	0.9988	0.1587	0.3093	0.3598	0.7514	0.1892	0.5081
Crackability	1.0000	0.9991	0.6663	0.9999	1.0000	0.9371	1.0000	0.9848	1.0000	0.9838

We conducted a two-tailed Kolmogorov-Smirnov Test, the null hypothesis being that the priming and non-priming password sets were from the same population concerning the metrics above. Since we could not find statistically significant differences between the priming and non-priming groups we concluded that priming did not have significant effects on our subjects within the respective categories. This enabled us to evaluate the effect of the type of study solely on the number of password sets we scored into the respective categories.

Table 7: Lab vs. Online (2-tailed Kolmogorov-Smirnov; P-Values).

	Null		**Single**		**Full**		**System**	
	real	study	real	study	real	study	real	study
Length	0.6878	0.7523	0.8868	0.5431	0.3741	0.4377	0.9972	0.9039
Shannon approx.	0.4551	0.7204	0.4890	0.6154	0.4624	0.3556	0.8727	0.8727
NIST	0.3550	0.9942	0.4304	0.4519	0.1509	0.8704	0.6734	0.6734
Digits	0.5154	0.9718	0.9996	0.4234	0.3458	0.4092	0.2770	0.9906
Upper Chars	0.9930	0.6332	0.6931	0.1710	0.8282	0.8236	0.5441	1.0000
Lower Chars	0.8680	0.4649	0.9444	0.2381	0.0435	0.2871	0.9972	0.8888
Special Chars	0.9598	0.9275	0.9119	0.9997	0.8645	1.0000	0.4435	0.4435
Crackability	0.6769	0.8034	0.9999	0.9315	0.9950	0.9863	0.7994	0.7994

We conducted a two-tailed Kolmogorov-Smirnov Test, the null hypothesis being that the password sets from the lab and the online participants in each category were from the same population concerning the metrics above. Since we could not find statistically significant differences between lab and online participants, we believe that our manual scoring was consistent irrespective of the type of study. This enabled us to evaluate the effect of the type of study solely on the number of password sets we scored into the respective categories.

Table 8: Entropy and Crackability Summaries for all Passwords of Participants

Real							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Shapiro-Wilk
Shannon approx.	47.26	55.57	62.52	63.64	69.79	99.79	$p < 0.0005$
NIST	18.00	25.50	30.75	29.77	33.50	42.00	$p < 0.0005$
Crackability	0.00%	0.0%	0.0%	16.82%	25.00%	100.00%	$p < 0.0005$
Study							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Shapiro-Wilk
Shannon approx.	47.26	56.02	63.88	64.87	71.45	102.80	$p < 0.0005$
NIST	24.00	30.75	32.63	32.86	34.50	42.00	$p < 0.0005$
Crackability	0.00%	0.0%	0.0%	15.47%	25.00%	100.00%	$p < 0.0005$

Table 9: Entropy and Crackability Summaries for all Passwords of Non-Participants

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Shannon approx.	47.45	56.56	63.51	64.43	71.28	103.10
NIST	24.00	30.75	32.50	32.64	34.50	42.00
Crackability	0.00%	0.00%	0.00%	17.87%	33.00%	100.00%

Table 10: Contingency - Table Self-Reported Different Password Behaviour in Our Study

	Self-Reporting (Fisher's Exact Test (alternative = greater) p<0.0005)		
Category	Different	Non-Different	Total by Category
Unhelpful	109	84	193
Helpful	148	304	452
Total by Self-Reporting	257	388	645

Table 11: Contingency Table - Study Completion by Scoring

	Study Completion (Fisher's Exact Test (alternative = two-tailed) p=0.9166)		
Category	Completed	Did-Not-Complete	Total by Category
Unhelpful	151	42	193
Helpful	356	96	452
Total by Completeness	507	138	645

Table 12: Contingency Table - Password Strategy by Scoring

	Password Strategy (Fisher's Exact Test (alternative=greater) p=0.01253)		
Category	Individual-Passwords	No-Individual-Passwords	Total by Category
Unhelpful	39	154	193
Helpful	57	395	452
Total by Strategy	96	549	645

Table 13: Contingency Table - Re-Login Rate by Scoring

	Re-Login (Fisher's Exact Test (alternative=two-tailed) p=0.6063)		
Category	Login-Success	Login-Failure	Total by Category
Unhelpful	165	28	193
Helpful	81	371	452
Total by Login Success	246	399	645

Table 14: Contingency Table - Scoring Results Online/Lab

	Scoring Results (Fisher's Exact Test (alternative=greater) p=0.0296)		
Category	Online	Lab	Total by Category
Unhelpful	181	12	193
Helpful	401	51	452
Total by Study Type	582	63	645