EAI Endorsed Transactions
on Security and Safety
Research Article EAI.EU

Human-Generated and Machine-Generated Ratings of Password Strength: What Do Users Trust More?

Saeed Ibrahim Alqahtani[1,*], Shujun Li[2,*], Haiyue Yuan[3], Patrice Rusconi[3]

[1]Taibah University, Madinah, Saudi Arabia
[2]University of Kent, Canterbury, UK
[3]University of Surrey, Guildford, UK

Abstract

Proactive password checkers have been widely used to persuade users to select stronger passwords by providing machine-generated strength ratings of passwords. If such ratings do not match human-generated ratings of human users, there can be a loss of trust in PPCs. In order to study the effectiveness of PPCs, it would be useful to investigate how human users perceive such machine- and human-generated ratings in terms of their trust, which has been rarely studied in the literature. To fill this gap, we report a large-scale crowdsourcing study with over 1,000 workers. The participants were asked to choose which of the two ratings they trusted more. The passwords were selected based on a survey of over 100 human password experts. The results revealed that participants exhibited four distinct behavioral patterns when the passwords were hidden, and many changed their behaviors significantly after the passwords were disclosed, suggesting their reported trust was influenced by their own judgments.

Received on 10 December 2019; accepted on 29 December 2019; published on 20 January 2020

Keywords: Password strength, password meter, user perception, trust, human-generated, machine-generated, ratings

doi:10.4108/eai.13-7-2018.162797

1. Introduction

Passwords have been dominating user authentication for more than half a century and many researchers believe that they will continue to represent a key part of user authentication in the foreseeable future, although many security and usability problems have been identified and a lot of new user authentication systems have been proposed over the years [1, 2].

Security problems of passwords are often caused by insecure behaviors of human users [1, 3, 4]. To help users create stronger passwords, proactive password checkers (PPCs), also called password (strength) meters by some researchers, have been widely used to encourage users to create stronger passwords by giving users real-time feedback on password strength [5]. Some PPCs have been found to be effective in leading to stronger passwords in some scenarios [6, 7].

All PPCs show a strength rating when a given password is entered by the user. Most widely used PPCs are based on simple heuristic approaches (i.e., password length and composition) to rate a password's strength [6]. The strength rating is either a categorical value such as "weak", "medium" or "strong" [8], or a numeric value such as an estimate of the password entropy [9], or the estimated guess number/time for the password being cracked [10].

More advanced PPCs are based on probabilistic or machine learning techniques in measuring a password's security against attacks. Examples include PPCs based on Markov models [11, 12], probabilistic context-free grammars [10, 13, 14] (PFCG), and neural networks [15]. Some researchers [16, 17] also looked at combining various heuristic and probabilistic approaches to design PPCs.

A special class of PPCs show the so-called peer pressure motivator (PPM), an estimate of a password's strength relative to the whole set of passwords chosen by all users [6, 18, 19]. Another recent development is

[*]Corresponding co-authors. Contact them via: Saeed Ibrahim Alqahtani – siqahtani@taibahu.edu.sa, Shujun Li – http://www.hooklee.com/.

a generalization of the concept of PPCs to a general-purpose Password Security Visualizer (PSV) reported in [20]. The PSV can be seen as a box of multiple parallel PPCs and other password information units, which are used together to inform users about different aspects of password security.

Strength ratings of passwords can be classified into machine-generated and human-generated ratings according to the extent of the direct human involvement in the strength evaluation, although the distinction between the two classifications can be more complicated. In this paper, we use the following definitions: 1) a machine-generated rating is an *automated* objective evaluation of password strength; 2) a human-generated rating is a subjective evaluation of password strength. Therefore, all strength ratings used in PPCs, regardless of whether they have a closer link to human perception of password strength, can be considered as machine-generated ratings.

Although there is some evidence of the usefulness of PPCs in motivating users to select stronger passwords, how human users perceive machine-generated ratings shown by PPCs and how such ratings influence their behaviors remain largely unknown. This is particularly problematic when the machine-generated ratings contradict the human-generated judgments of the users or what they heard from human experts. Such a contradiction can potentially cause a loss of user's trust on the PPC concerned, so understanding users' trust in such ratings could provide useful insights for the design and evaluation of PPCs and other password security tools. The determinants of trust in the human decision-making process have been studied in various contexts [21, 22]. However, to the best of our knowledge, they have not been systematically investigated in the context of password research except for some work [6, 23] that looked at user perception of password security.

To fill this gap, we conducted a user study on users' trust in human- and machine-generated ratings of password strength. Data collected from over 1,000 crowdsourcing workers revealed that: 1) user's own subjective perception of password strength could heavily influence their perceived trust in human- and machine-generated ratings of password strength; 2) there are different user-specific behavioral patterns in the reported trust in human- and machine-generated ratings of password strength; 3) users' trust in password ratings could be password-dependent.

The rest of the paper is organized as follows. Section 2 presents some related work. A detailed description of our user study is given in Section 3, followed by data analysis in Section 4 and more discussions of the above four main findings in Section 5. Sections 6 and 7 present limitations of this study and future work, respectively. The last section concludes the paper.

2. Related Work

2.1. Role of Trustworthiness

Trust in PPCs can be important in predicting to what extent PPCs will affect human users' decision making processes. Huang et al. [24] indicated that low perceived security may cause users to reject the use of IT system, while high perceived security may result in engaging insecure practices. This would imply that risk perception, which can be related to trust, can greatly affect users' decisions and behaviors. To the best of our knowledge, no research has investigated the effect of trust in human decision making in the context of password ratings.

However, in social sciences, trust is often defined as an individual's readiness for a vulnerable circumstance as a result of a positive expectation of others' actions [22], where "others" can be other people or automated systems. It is a general understanding that the interpretation of trust is context-dependent and research on it requires a multi-dimensional approach in the right context [25]. On the other hand, social psychological research has shown that individuals can infer a person's trustworthiness from a face in as fast as 100 ms [26]. In addition, trustworthiness is a morality-related personality trait, and previous work has shown that morality plays a primary role in social perception and social judgment [27–29].

Researchers have often attributed trust to three factors of a trustee: 1) expertise/ability (i.e., the degree to which a trustee is believed to be competent), 2) trustworthiness (i.e., the extent to which a trustee is believed to be cooperative and kind), 3) honesty (i.e., the degree to which a trustee is believed to have integrity) [22, 30]. Trust is also subject to other factors related to personality traits and conditional reasons [21, 22, 25]. Toma's study reported several cues linked to perceived trustworthiness in Facebook profiles [31].

2.2. User Perception of PPCs

Few studies have focused on users' perceptions of password strength, but some work looked at user perception around the use of PPCs. Ur et al. [6] conducted a large-scale user study with 2,931 crowdsourcing participants and 14 PPCs to collect participants' opinions on the influence of PPCs in creating passwords. The study showed that more stringent PPCs performed better in influencing users to create stronger passwords, but they also caused more complaints. Some participants expressed disagreements, surprises and even anger as a reaction to machine-generated ratings of some conservative PPCs. Participants' prior experiences usually had an impact on their impressions of password strength, particularly when they contravene their expectations.

A similar finding was also reported in [20], where a semi-structured interview study on password security showed that users trust PPCs in some situations especially when a PPC reports password strength that matches their expectations.

In another work conducted by Egelman et al. [7], two different PPCs were tested with real passwords used by participants in a lab setting and a field experiment. It was observed that the mere presence of a PPC could influence users to create stronger passwords. This work also led to another observation that participants *knowingly* chose weak passwords for unimportant accounts, which implies that users' subjective judgments on password strength did play a role in their decision-making process.

Sotirakopoulos et al. [18, 19] proposed a PPC design based on a peer pressure motivator (PPM). They conducted user studies to verify if PPM worked, and mixed results emerged: the PPM-based PPC did influence users positively compared with the case without any PPC, but it did not make any significant difference compared with other non-PPM PPCs. Egelman et al. got a similar finding in terms of its effectiveness [7].

Some research has focused on understanding the password creation process from the user's perspective. Ur et al. [32] conducted a qualitative user study aiming to understand common password patterns as well as investigating users' perception of password strength using a think-aloud and role-playing approach. They observed that although most users have a well-defined password creation strategy, some misconceptions about password strength (e.g., adding a digit or a special character to the end of a password makes it secure) could result in creating weak passwords. Furthermore, the misunderstanding of security advice can also cause misconceptions around password strength.

Other research has focused on understanding the users' perception of password strength suggesting that participants had serious misconceptions on how to make strong passwords, although in some other cases their perceptions of strong passwords matched password measures [23, 33]. A large variance in participants' understanding of password cracking methods was observed in [33], which suggests that human judgments on password strength are likely to be user-specific. Seitz and Hussmann [23] also suggested personalization as an effective motivator to secure behaviors.

2.3. Impact of Human/Machine-Generated Feedback on Decision Making

To the best of our knowledge, there is only limited research investigating the impact of human-generated ratings/feedback and machine-generated ratings on the human decision-making process, mostly in the health care, business and marketing literature. For instance, Lynn [34] found that a subjective message about the rise of venereal disease was trusted more than an objective message. Darley and Smith [35] investigated the effect of message board persuasion in terms of subjective messages and/or objective messages in marketing research. The use of objective messages was shown to be more effective than subjective messages, but the effectiveness of combining both types of messages did not differ significantly from the case of using one type of messages alone. They concluded that conditions and the context are more important factors to be considered. Research has also shown that people are averse to rely on algorithms and they opt for human judgments instead, a phenomenon known as algorithm aversion [36].

The most closely-related work in the cyber security area is from Chen et al. [37]. They conducted a user study to examine the impact of risk (negative) and safety (positive) information summaries of mobile apps on participants' decisions to install apps. Their results imply that developing a valid risk/safety index for mobile apps could potentially improve users' app-installation decisions, especially when this information is framed in terms of safety. Chong et. al. [38] managed to replicate the findings from Chen et al. when they study the influence of privacy priming and security framing on Android App selection. Yet, they proved there is a greater effect by scores framed as safety instead of risk.

3. User Study Design

As users' trust can be important in predicting to what extent people rely on PPCs to make decisions about password choices, our work focuses on the effect of trust on human decision-making in the context of password ratings (what is trusted more: machine-generated, human-generated or user's own judgment of password security). In other words, our work examines the question of what the most trusted evaluation source of password strength is: "machine-generated", or "human-generated" evaluation of password security, or their own judgment.

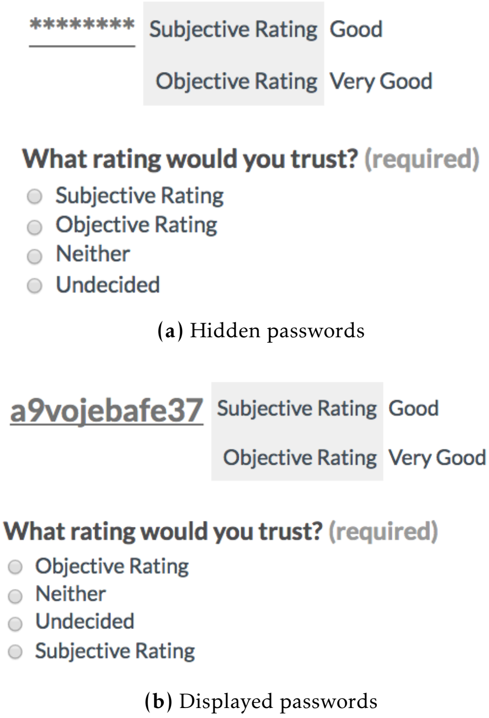
3.1. Hypotheses

The user study was designed to investigate users' trust in human-generated and machine-generated ratings of password strength. Our expectation was that users' trust (i.e., self-reported) in human-generated and machine-generated ratings of password strength would depend on several factors. Some of these factors can be linked directly to human factors such as users' experience and characteristics, or to non-human factors such as the password structure, and the type of feedback given (human/machine-generated).

We formulated four hypotheses: H1) users' own subjective judgment on password strength plays a significant role in users' self-reported trust in human-generated and machine-generated ratings; H2) users' self-reported trust in human-generated and machine-generated ratings is user specific; H3) users' self-reported trust in human-generated and machine-generated ratings is password-dependent; H4) some demographic factors play a significant role in users' self-reported trust in human-generated and machine-generated ratings.

3.2. Procedure

To test the above hypotheses, we conducted a within-subjects crowdsourcing user study. At the beginning of the user study, a brief overview was given to each participant explaining the meanings of human-generated and machine-generated "objective" ratings. The study was designed to be completed within 30 minutes. The study was structured in three sessions.[1] Participants' demographics including age, gender, and their computer skill levels were collected in the first session. In the next two sessions, six passwords with their realistic human-generated and machine-generated ratings (see Table 1) were presented to participants in two different conditions: 1) "hidden" – passwords shown as eight[2] asterisks; 2) "displayed" – passwords shown in clear (see Fig. 1). Passwords' ratings in both sessions are shown in clear. Having the two sessions allowed us to test Hypothesis H1 as disclosing passwords would give users new information to make their own judgment on the strength of each password, therefore influencing their perception of the human-generated and machine-generated ratings. For each password, participants were asked which password rating ("subjective" or "objective") they trusted more. They could also choose two other options ("neither" or "undecided") if they trusted neither or if they could not decide. Figure 1 shows example screenshots of questions we asked our participants.



******** Subjective Rating Good

Objective Rating Very Good

What rating would you trust? (required)

○ Subjective Rating
○ Objective Rating
○ Neither
○ Undecided

(**a**) Hidden passwords

a9vojebafe37 Subjective Rating Good

Objective Rating Very Good

What rating would you trust? (required)

○ Objective Rating
○ Neither
○ Undecided
○ Subjective Rating

(**b**) Displayed passwords

Figure 1. Screenshots of hidden and displayed password questions and rating options.

Table 1. Six passwords used in our experiments and their strength ratings, shown in 5-point scale (Very Weak, Weak, Medium, Good, Very Good).

ID	Password	machine -generated	human -generated
PW1	Q2W3E4R5	Weak	Very Weak
PW2	a9vojebafe37	Very Good	Good
PW3	St3v3J0b$Dropbox	Very Good	Weak
PW4	heart of darkness	Good	Weak
PW5	p@$$vv0rd	Medium	Very Weak
PW6	aAaAaAaA	Very Weak	Very Weak

All passwords with their ratings were presented to participants in the same order as shown in Table 1. We did not randomize the order for two reasons: 1) to get a complete randomization of password orders, this would have required a larger sample size of participants and additional costs could be avoided if the password order is fixed; 2) the results of a separate smaller user study indicated that the password order did not have any significant effect, supporting our decision to use a fixed order of passwords in the main experiment. This smaller user study involved 240 participants assigned to four groups (60 participants per group) and the six passwords with their ratings in each group were shown in different order. In Group 1, passwords were shown in a random order as shown in Table 1. In Group 2, passwords were arranged based on their strength (from the weakest to the strongest) and the order in Group 3 was reversed. In Group 4, passwords were arranged to have weak and strong passwords alternately. Then, we ran a multinomial regression analysis on the collected

[1] In our user study, we actually included three more sessions on privacy ratings of mobile apps for a different study. Participants were randomly divided into two groups, one was asked to do the password tasks first and the other the mobile app tasks first. Our analysis showed that the data from both groups were consistent so we used all participants for our data analysis reported in this paper.

[2] Fixing the length of hidden passwords is for controlling the effect of passwords length on participants' decisions on which rating is better than the other.

data which showed that there were no significant differences between the groups except between Groups 1 and 2.

We also ran another multinomial regression test comparing the data of the 240 participants collected from the smaller user study with the data of 240 participants who were randomly sampled from the main experiment where all passwords were shown on a fixed order. The multinomial regression results showed no significant differences between the two groups ($\chi^2 = 4.1955$, $p < 0.24111$, McFadden $R^2 = 0.000301$).

The six passwords and their human-generated ratings shown to participants were determined based on an online survey of over one hundred cyber security experts recruited from different channels (see Section 3.3 for more details). We used experts rather than average users because we believe that the former group has more knowledge needed to make appropriate judgments on password strength [39]. The machine-generated ratings of the passwords were obtained from the widely used PPC zxcvbn [40]. Participants were told that human-generated ratings were generated by a group of security experts while machine-generated "objective" ratings were generated by a computer algorithm developed by a group of security experts. Our aim was to use a neutral language to avoid wording that could influence participants' answers.

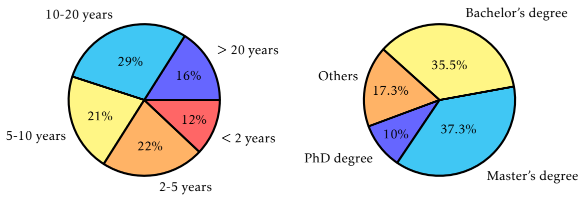
After each participant finished answering one question, a follow-up question was asked for the participant to select the reason of his/her answer out of a list of pre-defined options plus a free-format text area for additional information. We collected this information to have a better understanding of how participants made their decisions. To minimize the potential influence of the order of answer options in each question, we randomized the order of answer options so that each specific order got an equal number of participants.

As in typical crowdsourcing user studies, we did not collect any personal data or sensitive data. The user study was reviewed by our University Ethics Committee (UEC) and a favorable ethics opinion was secured before we started the user study.

3.3. Passwords and Ratings Determination

We decided to use only six passwords to keep participants workload light. To reduce password selection biases, the six passwords were selected from a larger set of 21 passwords, which were studied in an online survey on those passwords' strength and categorization. The 21 passwords were chosen from a wide range of password categories and complexity (e.g., common passwords, alternation to common passwords, personal-based passwords, and random passwords). The survey was published through a number of cyber security channels (the Openwall password (http://openwall.com/lists/passwords) and NCC Group mailing lists, the cyber security community on Freenode (http://irc.freenode.net) and Twitter). It also collected self-reported basic demographic information to exclude non-expert participants.

In total, 110 (self-recognized) password experts took our survey. Figure 2 shows the level of experience and education of all participants. Most of the 21 passwords were rated by all experts, but password categories received fewer answers, suggesting that some experts had difficulties in making decisions on the best category for some passwords. Interestingly, the passwords' human-generated ratings collected from human experts had a general tendency to be more conservative than the corresponding machine-generated ratings given by a state-of-the-art PPC zxcvbn [40]. This could represent a bias in the machine-generated and human-generated password strength as the difference between human-generated and machine-generated ratings used could influence users' perceived trust in such ratings. However, we decided to trade off this bias against the ecological validity of our password samples. Further research on experts' perception of password strength is required to clarify how human-generated password ratings could be collected and used in a different fashion.



(a) Level of Experience

(b) Level of Education

Figure 2. Demographics of experts recruited.

From the 21 passwords we selected six based on two criteria: 1) ensuring distinct levels of password structure and complexity, and thus different levels of guessing effort so that we could test hypothesis H3; 2) ensuring a relatively wide range of human-generated and machine-generated ratings while keeping the participants' overall task light to avoid biases due to fatigue.

According to the selection criteria, the six passwords shown in Table 1 were selected from three major password categories commonly used in real-world scenarios: 1) "word password": a string that consists of one or more common words with or without character transformation; 2) "non-word password": a string that is based on keyboard patterns or repeated patterns and do not contains any words; 3) "mixture password": a more complicated string that can be segmented into different segments of different category. We also selected the six passwords so that they all had distinct human/machine-generated rating combinations, which also helped us conduct the analysis on password dependency of users' reported trust in human-generated and machine-generated ratings.

The selected six passwords shown in Table 1 represent three major categories of passwords human users often use in real-world scenarios: 1) dictionary-based passwords (i.e., "p@$$vv0rd") and pass-phases (i.e., "heart of darkness"); 2) passwords formed based on keyboard patterns (i.e., "Q2W3E4R5") and a repeated structure ("aAaAaAaA"); 3) passwords formed based on hybrid rules (i.e., "St3v3J0b$Dropbox") and random characters (i.e., "a9vojebafe37").

In terms of human-generated ratings, we decided to use the median of all experts' reported ratings as the human-generated ratings used in our user study. In the case that the median rating lay exactly between two ratings (e.g., for the median value 0.5, both 0 ("very weak") and 1 ("weak") are equally distant from the median value), we selected the rating receiving more votes. We managed to decide the ratings of all the six passwords following these two rules.

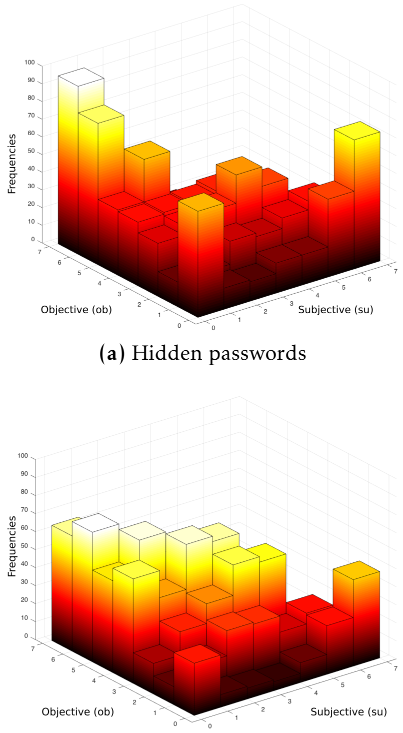
3.4. Participant Recruitment

The crowdsourcing platform CrowdFlower[3] was used to recruit participants. Each participant was rewarded $0.6 for the whole study (including the mobile app tasks not covered in this paper). We decided the price of $0.6 based on similar tasks on the same platform to avoid under- or over-motivate participation. Note that on crowdsourcing platforms rewards for such micro tasks are mostly very cheap.

We recruited only the most trusted workers rated by CrowdFlower (i.e., the "Level 3" workers) to maximize data quality. We also split the whole user study into a number of parts and ran them on different days and at different times to recruit participants from a wider spectrum in terms of geo-locations and working times.

4. Results

In total 1,100 participants took part in our user study. 23 participants were excluded because they took the experiment more than once. Therefore, the data we analyzed were based on $1,100 - 23 = 1,077$ participants from 68 countries. The male-female ratio was 66% to 34%. The participants had a reasonably wide age range: 26-35 (39.4%), 18-25 (30.6%), 36-45 (20.8%), over 45 (8.6%), and below 18 (0.6%). Due to the relatively small number of participants in the last two age groups, we re-grouped participants to have a more even distribution: 25 or below (336 participants), 26-35 (424 participants), 36 or older (317 participants).



(a) Hidden passwords

(b) Displayed passwords

Figure 3. 2-D histograms of participant's collective behaviors for choosing "su" and "ob".

Next, we report the main results of our user study. For the four answer options for each password, we define 2-character short names: "su" for "human-generated rating", "ob" for "machine-generated rating", "ne" for "neither", and "ud" for "undecided".

4.1. Behavioral Analysis: Password Condition

First, we report the impact of the password display condition ("Hidden" or "Displayed", as a binary variable) on participants' self-reported trust in human-generated and machine-generated ratings. Here, the dependent variable is the 4-valued answer of each participant (i.e., the self-reported trust). Among the four values, we are more interested in "su" and "ob", but also examine "ne" and "ud" since they can reveal how participants felt about the shown human-generated and machine-generated ratings.

In order to visualize all participants' collective behavior, we produced 2-D histograms of participants' behavioral patterns for both "Hidden" and "Displayed" conditions as shown in Fig. 3, where the 2-D bin at position (i, j) indicates the number of participants who chose "su" i times and "ob" j times ($0 \leq i + j \leq 6$). The maximum value of $i + j$ is 6 because the sum cannot

[3] Renamed to "Figure Eight" (https://www.figure-eight.com/) in 2018.

exceed the number of passwords shown to participants, which is 6.

From Fig. 3, we can visually observe that the participants' collective behavior was affected by the password display condition. When the passwords were hidden, most participants tended to follow one of four typical behaviors (i.e., four peaks in the 2-D histogram – see Section 4.2). However, once the passwords were disclosed, the majority of participants changed their choices for at least one password so that the shape of the 2-D histogram changed drastically with much less clear peaks.

For participants who reported full trust in either machine-generated or human-generated ratings (95 and 73), 39% of them (30 out of 73 participants who had full trust in human-generated rating and 35 out of 95 participants who fully trusted machine-generated rating) changed their reported trust for at least one password, leading to a much flatter histogram. It deserves noting that around 3% (32/1077) of the participants reported trust in neither human-generated nor machine-generated ratings in both conditions, suggesting that some human users may have an intrinsic disbelief on ratings given by others (regardless of whether the sources are machines or other people).

Although the differences between the two 2-D histograms are clearly visible, we applied Stuart-Maxwell χ^2 tests (with a degree of freedom of 3 because the dependent variable has 4 values) and a multinomial regression to test if the differences were statistically significant.

The Stuart-Maxwell χ^2 tests showed that the observed difference was indeed statistically significant as seen in Table 2.[4] However, the multinomial regression results in Table 3 show a low McFadden's R^2 value, suggesting that the condition as a binary variable did not have a good predictive power so the differences were better tested by other statistical tests (i.e., Stuart-Maxwell χ^2 test). In the table, each row represents a linear prediction model $\ln\left(\frac{p(y)}{p(\text{ob})}\right) = \beta_{y0} + \beta_{y1} \times x$, where the predictor variable x is the display condition (1 = display, 0 = hidden), the predicted variable is $\ln\left(\frac{p(y)}{p(\text{ob})}\right)$, and $y \in \{\text{su}, \text{ne}, \text{ud}\}$.

4.2. Behavioral Analysis: Behavioral Pattern

As we mentioned before, in Fig. 3(a) we can observe four peaks, each referring to a different behavioral pattern. This seems to suggest that each participant had some intrinsic behavioral style that could influence their self-reported trust in human-generated and machine-generated ratings. Therefore, by knowing which behavioral style a person had, his/her trust in human-generated and machine-generated password ratings could be predicted which allowed us to test H2. Therefore, we ran a k-means algorithm to cluster all participants of our user study into four behavioral clusters: P1 (human-generated rating believer, 220 participants, centre={5,3}), P2 (machine-generated rating believer, 410 participants, centre={0.5,4.6}), P3 (balanced believer, 242 participants, centre={2.9,2.2}) and P4 (disbeliever, 205 participants, centre={0.7,0.9}.

Table 2. Results of the Stuart-Maxwell χ^2 tests for analyzing participants' self-reported trust in machine-generated and human-generated ratings.

Distributions Compared	χ^2	p-value
su (Hidden) vs. su (Displayed)	46.079	2.86×10^{-8}
ob (Hidden) vs. ob (Displayed)	98.349	2.20×10^{-16}
su (Hidden) vs. ob (Hidden)	49.051	7.28×10^{-9}
su (Displayed) vs. ob (Displayed)	120.29	$< \varepsilon$

Table 3. Results of the multinomial logistic regressions conducted on the password display condition as the predictor of participants' self-reported trust.

Predictor	Option	b	SE	p-value	OR
Displayed	su	-0.15	0.04	1.9×10^{-4}	0.858
Displayed	ne	-0.01	0.06	0.92	0.995
Displayed	ud	-0.49	0.06	$< \varepsilon$	0.614

$\chi^2 = 78.841$ ($p < \varepsilon$), McFadden R^2: 0.002. The baseline of the independent variable (password display condition) is "Hidden".

We conducted another multinomial logistic regression using the behavioral cluster of each participant obtained from the k-means clustering. This regression evaluates whether the behavioral cluster label is a good predictor of participants' perceived trust. The results are depicted in Table 4, which indicate that the overall effect is statistically significant with mostly significant odds ratios. The results show that human-generated rating believers (P1) and balanced believers (P3) are more likely to select human-generated ratings over machine-generated ratings compared to the disbelievers (P4). The odds ratio of (P1) shows that human-generated rating believers are predicted to select human-generated ratings over machine-generated ratings more than those who belong to the other behavioral styles.

The above analysis may be seen as circular reasoning as the personality labels are obtained from the data and then used to predict the data. To further validate whether the personality labels obtained from running the k-means clustering are reliable, we ran a new

[4] For p-value, "$< \varepsilon$" means that the exact p-value could not be obtained but it drops below the precision limit (which is 2.22×10^{-16} for R, the language we used for statistical tests). The same notation will be used for other tables throughout this paper.

Table 4. Results of the multinomial logistic regression conducted on the behavioral pattern as the predictor of participants' self-reported trust.

Predictor	Option	b	SE	p-value	OR
P1	su	1.69	0.08	$< \varepsilon$	5.392
P1	ne	-1.35	0.12	$< \varepsilon$	0.259
P1	ud	-1.19	0.1	$< \varepsilon$	0.305
P2	su	-1.5	0.08	$< \varepsilon$	0.222
P2	ne	-1.94	0.07	$< \varepsilon$	0.143
P2	ud	-2.72	0.08	$< \varepsilon$	0.066
P3	su	0.31	0.08	3.54×10^{-5}	1.367
P3	ne	-1.55	0.09	$< \varepsilon$	0.212
P3	ud	-2.04	0.09	$< \varepsilon$	0.131

$\chi^2 = 4576.1$ ($p < \varepsilon$), McFadden R^2: 0.1428. The baseline of the independent variable is "P4".

analysis where we split the data into two non-overlapping subsets, as can be see at Table 5. Each subset contained users responses on a different subset of three passwords. Then, we ran k-means clustering algorithm on each data subset to derive the personality label for each participant and then used the label as an independent variable to predict the reported trust in the other subset. Next, we conducted a multinomial regression on each data subset. The results showed that the odds ratios observed in the new analysis were aligned with the finding in the first analysis, indicating that most users behave consistently for different passwords in how they reported their trust in human-generated and machine-generated ratings.

We were also interested in the behavioral changes of participants with different behavioral patterns when the password display condition changed from "Hidden" to "Displayed". Figure 4 shows a comparison between the distribution of users' responses in terms of their choices on "su" and "ob" for the four behavioral patterns. There are four green sub-figures, each refers to a particular behavioral group style (P1, P2, P3, or P4) and highlights the distribution of users' responses when passwords were hidden. The number of participants in a particular behavioral group style is shown at the top of the green sub-figure. The four yellow sub-figures highlight how users with a particular behavioral style changed their behaviors when passwords were displayed. At the top of each sub-figure, the number of participants who did not change their behavior is highlighted in dark grey while the total number of participants who completely shifted to another behavioral style is highlighted in dark red. The number of participants in each of the other behavioral styles was highlighted in light red.

As a whole, 54% of participants (581/1077) changed their reported trust. More than half (126/220, 57%) of human-generated rating believers (P1) changed their reported trust for at least one password (see Fig. 4a). Most of them (19%) had an extreme shift towards trusting machine-generated ratings while a few of them (5%) shifted to disbelievers. However, machine-generated ratings believers (P2) seemed to have a stronger view as they shifted only slightly towards balanced believers or disbelievers, and only 8 (2%) completely changed their positions (see Fig. 4(b)). For balanced believers (P4), the behavioral change had a wider distribution, whereby 87 (36%) participants migrated to trust machine-generated ratings more (see Fig. 4(c)). Finally, P4 had the similar behavioral distribution as P3 but with low level of trust in human-generated ratings (see Fig. 4(d)).

Table 5. Results of a multinomial logistic regression conducted on two password data subsets as the predictor of participants' self-reported trust.

(**a**) Model 1 based on Dataset 1

Predictor	b	p-value	OR
P1:ot vs ob	-1.210	2.32×10^{-10}	0.298
P1:sb vs ob	1.619	$< \varepsilon$	5.050
P1:ud vs ob	-0.305	0.024	0.737
P2:ot vs ob	-1.198	$< \varepsilon$	0.302
P2:sb vs ob	-0.807	$< \varepsilon$	0.446
P2:ud vs ob	-1.683	$< \varepsilon$	0.186
P3:ot vs ob	-0.996	1.13×10^{-12}	0.369
P3:sb vs ob	0.508	2.41×10^{-7}	1.663
P3:ud vs ob	-0.884	1.76×10^{-12}	0.413

[a] $\chi^2 = 1357.5$ ($p < \varepsilon$), McFadden R^2: 0.082. The dataset includes users responses on PW1, PW3, and PW5.

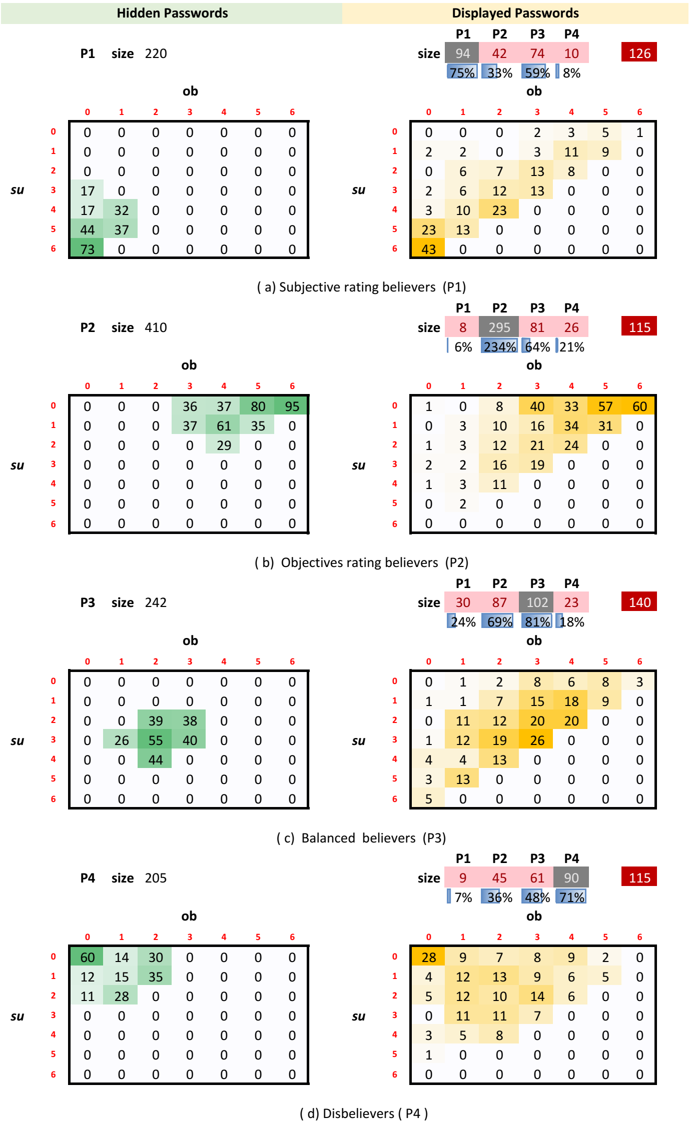
[b] The baseline is "P4". The reference of password rating is "ob".

(**b**) Model 2 based on Dataset 2

Predictor	b	p-value	OR
P1:ot vs ob	-1.204	1.41×10^{-6}	0.300
P1:sb vs ob	1.044	$< \varepsilon$	2.841
P1:ud vs ob	-2.280	1.54×10^{-12}	0.102
P2:ot vs ob	-1.034	$< \varepsilon$	0.356
P2:sb vs ob	-1.449	$< \varepsilon$	0.235
P2:ud vs ob	-2.011	$< \varepsilon$	0.134
P3:ot vs ob	-0.851	3.13×10^{-12}	0.427
P3:sb vs ob	-0.325	2.14×10^{-4}	0.722
P3:ud vs ob	-1.809	$< \varepsilon$	0.164

[a] $\chi^2 = 1217.5$ ($p < \varepsilon$), McFadden R^2: 0.078. The dataset includes users responses on PW2, PW4, and PW6.

[b] The baseline is "P4". The reference of password rating is "ob".



P1 size 220

ob / su

su \ ob	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	17	0	0	0	0	0	0
4	17	32	0	0	0	0	0
5	44	37	0	0	0	0	0
6	73	0	0	0	0	0	0

	P1	P2	P3	P4	
size	94	42	74	10	126
	75%	33%	59%	8%	

su \ ob	0	1	2	3	4	5	6
0	0	0	0	2	3	5	1
1	2	2	0	3	11	9	0
2	0	6	7	13	8	0	0
3	2	6	12	13	0	0	0
4	3	10	23	0	0	0	0
5	23	13	0	0	0	0	0
6	43	0	0	0	0	0	0

(a) Subjective rating believers (P1)

P2 size 410

su \ ob	0	1	2	3	4	5	6
0	0	0	0	36	37	80	95
1	0	0	0	37	61	35	0
2	0	0	0	0	29	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

	P1	P2	P3	P4	
size	8	295	81	26	115
	6%	234%	64%	21%	

su \ ob	0	1	2	3	4	5	6
0	1	0	8	40	33	57	60
1	0	3	10	16	34	31	0
2	1	3	12	21	24	0	0
3	2	2	16	19	0	0	0
4	1	3	11	0	0	0	0
5	0	2	0	0	0	0	0
6	0	0	0	0	0	0	0

(b) Objectives rating believers (P2)

P3 size 242

su \ ob	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	39	38	0	0	0
3	0	26	55	40	0	0	0
4	0	0	44	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

	P1	P2	P3	P4	
size	30	87	102	23	140
	24%	69%	81%	18%	

su \ ob	0	1	2	3	4	5	6
0	0	1	2	8	6	8	3
1	1	1	7	15	18	9	0
2	0	11	12	20	20	0	0
3	1	12	19	26	0	0	0
4	4	4	13	0	0	0	0
5	3	13	0	0	0	0	0
6	5	0	0	0	0	0	0

(c) Balanced believers (P3)

P4 size 205

su \ ob	0	1	2	3	4	5	6
0	60	14	30	0	0	0	0
1	12	15	35	0	0	0	0
2	11	28	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

	P1	P2	P3	P4	
size	9	45	61	90	115
	7%	36%	48%	71%	

su \ ob	0	1	2	3	4	5	6
0	28	9	7	8	9	2	0
1	4	12	13	9	6	5	0
2	5	12	10	14	6	0	0
3	0	11	11	7	0	0	0
4	3	5	8	0	0	0	0
5	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0

(d) Disbelievers (P4)

Figure 4. 2-D distribution of participants' responses according to behavioral patterns and password display conditions.

4.3. Behavioral Analysis: Demographic Factors

We conducted an additional multinomial logistic regression to test the possible impact of demographic factors including gender, age, and skill level on participants' trust. The results showed that the effect was not significant ($\chi^2 = 321.77$, $p < \varepsilon$, McFadden $R^2 = 0.01$, odds ratios are mostly not far from 1).

4.4. Behavioral Analysis: Password Dependencies

We also conducted a password-level analysis to see if participants' behaviors depended on an individual password. Since the human-generated and machine-generated ratings (and their difference) were dependent on passwords, what we considered were actually both the passwords and their ratings. If and how we can separate these two aspects remains an open question for future studies (which most likely would require fictitious ratings with the additional disadvantage that they can be easily detected by participants). As we aimed at studying the effect of password, password dependency would refer to not only the effect of password but to its own human-generated and machine-generated rating. Since the value of password strength ratings varies according to the concerned password, we cannot study the effect of password structure alone.

Table 6. Results of the Stuart-Maxwell χ^2 tests on participants' perception of human-generated and machine-generated password ratings for different passwords ("hidden" to "displayed").

Predictor	χ^2	p-value
PW1(H) vs PW1(D)	54.9	7.2×10^{-12}
PW2(H) vs PW2(D)	37.3	4.0×10^{-8}
PW3(H) vs PW3(D)	293.6	$< \varepsilon$
PW4(H) vs PW4(D)	65.5	4.0×10^{-14}
PW5(H) vs PW5(D)	32.8	3.5×10^{-7}
PW6(H) vs PW6(D)	12.1	0.007

D: the displayed password condition; H: the hidden password condition.

We used a Stuart-Maxwell χ^2 test to see if the distribution of responses' shifted significantly when the password changed, allowing us to test Hypothesis H3. In Table 6, the results showed significant differences for all passwords when the password display condition was changed from "Hidden" to "Displayed".

Table 7 shows a comparison between different password pairs when passwords were displayed. A number of Stuart-Maxwell χ^2 tests were used for testing homogeneity for the four rating options ("su","ob","ne" and "ud"). The results also showed significant differences between different password pairs, suggesting that users' trust and decision-making were password dependent.

We also conducted a multinomial regression to see the predictive effect of the password on selecting either human-generated or machine-generated ratings. We chose PW6 ("aAaAaAaA") as the baseline since among all passwords it has the simplest structure and it had the least influence on users' trust in human-generated and machine-generated ratings. The multinomial regression results showed an overall significant difference between different passwords ($\chi^2 = 1072.2$, $p < \varepsilon$, McFadden $R^2 = 0.00033$).

Table 8 shows significant differences for most of passwords except PW1 ("Q2W3E4R5") and PW4 ("heart of darkness") in terms of users' reported trust in human-generated ratings and machine-generated ratings. The results showed that participants were more

Table 7. Results of the Stuart-Maxwell χ^2 tests on participants' perception of human-generated and machine-generated password ratings for different displayed password pairs. The p-values for all cases are $< \varepsilon$ except for two cases: "PW1 vs PW5" ($p = 1.3 \times 10^{-10}$) and "PW2 vs PW5" ($p = 2.6 \times 10^{-8}$).

Predictor	χ^2	Predictor	χ^2
PW1 vs PW2	130.3	PW1 vs PW3	242.5
PW1 vs PW4	125.8	PW1 vs PW5	49.0
PW1 vs PW6	132.4	PW2 vs PW3	233.1
PW2 vs PW4	101.1	PW2 vs PW5	38.2
PW2 vs PW6	328.0	PW3 vs PW4	390.3
PW3 vs PW5	205.8	PW3 vs PW6	361.4
PW4 vs PW5	103.4	PW4 vs PW6	299.1
PW5 vs PW6	227.6		

likely to select machine-generated ratings over human-generated ratings for PW2, PW3, PW5, in relation to PW6. As a whole, we can conclude that users' reported trust was password dependent.

Table 8. Results of multinomial logistic regressions conducted on the displayed passwords as the predictor of participants' self-reported trust.

	Predictor	b	p-value	OR
PW1	su vs ob	-0.052	0.514	0.950
	ne vs ob	-0.879	$< \varepsilon$	0.415
	ud vs ob	-1.378	$< \varepsilon$	0.252
PW2	su vs ob	-0.209	0.007	0.811
	ne vs ob	-2.144	$< \varepsilon$	0.117
	ud vs ob	-1.676	$< \varepsilon$	0.187
PW3	su vs ob	-0.485	7.79×10^{-10}	0.616
	ne vs ob	-1.893	$< \varepsilon$	0.151
	ud vs ob	-1.616	$< \varepsilon$	0.199
PW4	su vs ob	0.148	0.056	1.159
	ne vs ob	-1.613	$< \varepsilon$	0.199
	ud vs ob	-1.283	$< \varepsilon$	0.277
PW5	su vs ob	-0.252	0.001	0.777
	ne vs ob	-1.342	$< \varepsilon$	0.261
	ud vs ob	-1.421	$< \varepsilon$	0.242

$\chi^2 = 1072.2$ ($p < \varepsilon$), McFadden R^2: 0.033459. The baseline is PW6. The reference of password rating is "ob".

Furthermore, as shown by participants' actual responses on each rating option for all six passwords, we found that participants' reported trust was mostly not password dependent when the passwords were hidden, except for PW6. Participants had a nearly uniform response among the four possible answers for PW6. This can be explained in light of the fact that PW6 had the same human-generated and machine-generated ratings, so participants made a random guess.

However, this was not the case when passwords were displayed. It is obvious that displaying passwords played a role in users' perception of trust, which led to very different responses among all participants (likely driven by their different behavioral styles). When the password complexity was not obvious to users such as for PW3 ("St3v3J0b$Dropbox") and PW5 ("p@$$vv0rd"), machine-generated ratings were more likely to be selected, unlike the case of PW4. Interestingly, users' perception of trust for PW6 did not vary much, which could be attributed to the reason explained above.

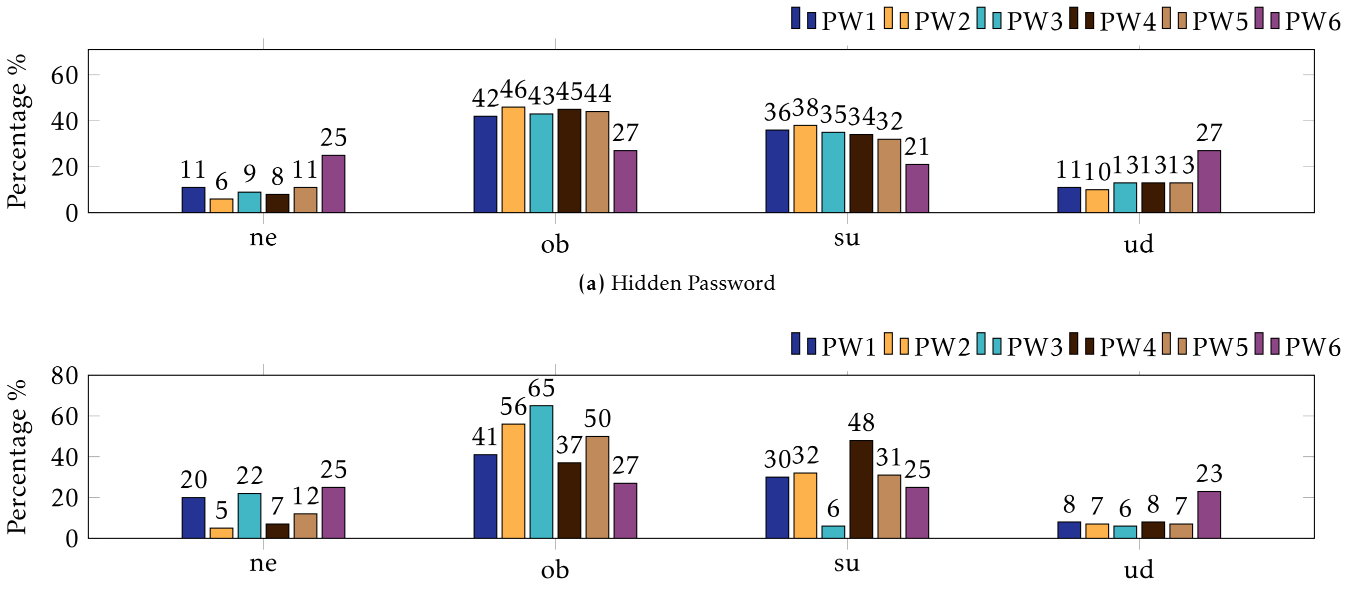
The above results can also be seen visually in Figure 5, which shows percentages of participants' answers for each of the four options and for all the five passwords. Figure 5(a) shows that participants had a convergence of views in terms of trust when passwords were hidden. However, this is not the case when passwords became clear as shown in Fig. 5(b). It is obvious that displaying passwords played a role in the users' perception of trust. Both the above regression tests and visual inspection of Fig. 5(b) lead to the same observation that the five passwords can be put into two groups: 1) PW1 and PW2 (more participants chose to trust human-generated ratings); 2) PW3, PW4 and PW5 (more participants chose to trust machine-generated ratings).

4.5. Users' Self-Reported Reasons of Trust Choices

We also collected users' self-reported reasons behind their choices of trust (see Appendix A for a list of predefined reasons depending on each user's choice of rating). In total, 16% of participants reported that machine-generated ratings were trusted more readily as they were generated by automated algorithms, which can detect hidden things better than humans. 40% of participants selected machine-generated or human-generated ratings as they matched their expectations while 10% of participants selected "neither" as none of the ratings matched their expectations. 15% of those who selected human-generated ratings were influenced by their desire to be on the safer side, mainly because human-generated ratings were more conservative than machine-generated ratings. Interestingly, none of those who selected human-generated ratings reported loss of trust in machine-generated ratings, while some of those who more readily trusted machine-generated ratings reported a loss of trust in human-generated ratings.

5. Discussion

Our work compares users' trust of password ratings given by two rating sources (PPCs and human password experts) thus providing an original contribution to the literature, in particular with respect to Ur et al.'s work [33]. This section expands on the interpretation of the



(**a**) Hidden Password

(**b**) Displayed Password

Figure 5. Percentages of participants' answers for all six passwords when they were hidden and displayed.

results, focusing on the four hypotheses we previously explained in Section 3.1.

5.1. Contextual Effects

The password display condition had an influence on users' decision making. Our findings showed a significant shift of the participants' collective behavior when password display condition changed, which confirms our first hypothesis (H1). This would also suggest that generalizing the results to different contexts (i.e., the context of mobile apps' ratings) might also be possible.

In addition, participants tended to trust machine-generated ratings more than human-generated ratings when the passwords were hidden. The results also showed that users' own subjective judgments on password strength played an active role in their trust in both password ratings. This was more obvious when passwords were displayed as the participants were supplied with more information.

Many participants preferred their own judgments of password strength when the passwords were disclosed. This shows that some people would apply their rational thinking to evaluate the trustworthiness of a trustee (i.e., source of password ratings or password rating) before placing trust.

We also observed that some participants were risk-averse as they showed willingness to select the rating that matched their expectations while some others had a higher tendency to trust a trustee even when there was not enough information. These behaviors can be attributed to the strong impact of people' behavioral patterns on their trust perception.

5.2. Behavioral Patterns and Their Effects

Our experiment revealed the existence of different behavioral patterns that had a significant impact on users' reported trust, providing support for hypothesis H2. Some participants appeared to be conservative and cautious when it came to trust a particular type of password ratings. This is reflected in some participants' tendency towards selecting the lower, more conservative ratings (i.e., choosing "Very Weak" instead of "Weak") to be on the safer side. This finding suggests that password rating can also have an influence on perceived trust.

Some other behavioral patterns can be associated with trust bias. Participants who had an extreme trust in either human-generated or machine-generated ratings can be considered relatively risk-seeker. In contrast, some participants seemed to be risk-averse as they avoided trusting any rating in most cases, regardless of the specific situation.

The effect of the behavioral patterns becomes less obvious with displayed passwords. This may be attributed to the fact that many participants based their reported trust on their own subjective judgments on the password strength, i.e., their behavioral style played a less important role than their judgment).

5.3. Impact of Individual Password on User's Trust

Our results showed that the password complexity and structure had a clear-cut influence on users' reported trust (hypothesis H3). This was observable when different displayed passwords received different trust responses.

It seems that participants could easily make a judgment when passwords had a simple structure such as PW4 (“`heart of darkness`”), leading to select human-generated ratings as they almost assigned lower ratings to passwords. For more complicated passwords, i.e., PW1 (“`Q2W3E4R5`”), PW2 (“`a9vojebafe37`”), PW3 (“`St3v3J0b$Dropbox`”) and PW5 (“`p@$$vv0rd`”), machine-generated ratings were more likely to be selected by human participants, which could be considered as indirect evidence that they had a good level of trust in PPCs. For PW3 (“`St3v3J0b$Dropbox`”), which had a complicated structure and a large difference between its human-generated and machine-generated ratings, users reported to trust the machine-generated rating more than the human-generated rating. These findings suggest that the use of PPCs is useful, as long as they report reasonable ratings.

Participants’ reported trust in human-generated and machine-generated ratings for PW6 (“`aAaAaAaA`”), which has identical human-generated and machine-generated ratings, shows that a significant portion of (25%) participants had a good level of trust in experts’ judgments (comparable with participants who trusted machine-generated ratings more readily – 27%). This can be a sign of the usefulness of having real experts’ password strength ratings in PPCs.

5.4. Demographic Factors

We did not observe any significant influence of demographic factors (gender, age and skill level) on users’ trust in human-generated and machine-generated ratings. This did not confirm Hypothesis H4. For gender and skills, this may be linked to the effect of an unbalanced sample, which is one of the limitations of this study.

5.5. Engagement of Participants

We would like to point out that many participants changed their responses after seeing the passwords, which suggests that they were actively engaged in the user study. This could be seen as indirect evidence that the observed changes are a reflection of behaviors that could be observed also in real-world settings. We acknowledge the natural limitations of using crowdsourcing workers for conducting user studies especially on the quality of data collected, but the nature and some design elements of our study (simple tasks that crowdsourcing workers could be motivated to engage in without providing random responses) gave us some level of confidence on the results we reported in this paper. In the future, we hope to conduct an even larger scale study with more passwords and more crowdsourcing workers and also a medium-scaled lab-based study to further validate the results in this paper.

6. Limitations

Our choice of passwords and their human-generated and machine-generated ratings impose some limitations. Although we attempted to select representative passwords for the experiment, the number of passwords we used (6) is small. Using a larger set of passwords would help reduce password selection biases and produce more convincing evidence of the findings reported. In addition, users’ trust can be influenced by many factors [21, 22], which are not easy to control in a single user study. These factors can include perception of password strength ratings, password composition, demographic factors, users’ brand loyalty (i.e., people may trust the rating obtained from a specific well-known password metric or from a trusted security community), context of use, etc.

We mentioned above that the unbalanced password rating differences was also expected to influence the results. However, this issue seems to be difficult to address in future research. The use of more balanced human-generated and machine-generated ratings might not be possible without using fictitious password human-generated and machine-generated ratings. If this is the case, it will contradict our goal to use realistic passwords and maintain ecological validity.

Furthermore, the password strength ratings themselves might influence users’ perception. Therefore, we have to consider this factor to analyze users’ behaviors accurately. This requires determining an accurate evaluation of password strength. This may be hard to do for both human-generated and machine-generated ratings since machine-generated ratings are not well defined, and “ground-truth” human-generated ratings need to consider opinions of a large number of security experts.

As we mentioned above, the behavioral analysis of the password display condition implies that it is more likely that many (if not most) participants were well engaged with the task. One question that needs to be addressed, however, is whether participants who did not change their answers for both password sessions were actually engaged to show their their genuine behaviors. Although there is a possibility of cheating, there is no clue about the actual number of cheating behaviors or misunderstandings of our questions. This intrinsic problem could be attributed to the use of a crowdsourcing platform, and future research is needed to see if this issue can be studied with more evidence about the level of engagement of each individual participant.

7. Future Work

Our study produced some surprising results. Particularly, the lack of observed effects of demographic factors was unexpected. While we could speculate about a number of possible explanations, the results imply that users' perceived trust and their knowledge on passwords are more complicated than we expected. The results may also be related to the limitations of the crowdsourcing method itself, whereby the demographic information provided by participants may contain much more noise than other more controlled settings. The influence of demographic factors requires further investigation.

Although the reported work is about password strength ratings only, we also conducted a parallel study on human-generated and machine-generated privacy ratings of mobile apps. Our results showed that participants' collective behaviors differed from those in the password case, which led us to believe that the application context also matters.

Given the observation that a significant number of participants chose to trust human-generated ratings only, introducing semi-subjective ratings into PPCs may be useful at least for some users. Such human-generated ratings can be collected based on a human-in-the-loop approach where experts, assuming that they are trustworthy, are encouraged to submit their own subjective ratings when they disagree with the password meter's machine-generated ratings. The extracted passwords features with their human-generated ratings can be then used as useful training data to simulate experts' opinions on unknown passwords' strength. This approach can then help improve password meters by fixing errors and producing more reliable machine-generated ratings. Human-generated ratings can be pooled in a way to keep only reliable ones from real experts. One way to determine the expertise of new users can be done through getting an acknowledgment from other known or pre-acknowledged experts or legal authorities. The approach can be used in combination with machine-generated ratings to enhance user choices of password. We would not underestimate the difficulties of actually implementing the human-in-the-loop idea, but this can lead to interesting future research on an aspect the password research community has not yet explored.

As a side outcome, this work also reports a study (the first according to the best of our knowledge) on human experts' strength ratings on 21 passwords and the categories they belong to. The study produced unexpected results that experts may be more conservative in rating passwords than PPCs are. We plan to further investigate this observed phenomenon in future work.

As a whole, more future research is needed to accumulate more evidence on how users perceive password strength ratings and PPCs and how they choose what to trust. Particularly, considering the limitations of any crowdsourcing based studies, we plan to conduct more crowdsourcing-based studies and also traditional lab-based studies to further validate the results we reported in this paper. Such studies will help the design and deployment of password checkers, passwords policies, and password educational tools.

In future work, the difference between human-generated and machine-generated ratings in users' trust in machine-generated and human-generated ratings requires some special handling of the ratings used (e.g., we may have to use false human-generated ratings to cover positive differences between human-generated and machine-generated ratings).

8. Conclusion

To the best of our knowledge, this paper reports the first study comparing users' perceived trust on password strength ratings given by automated algorithms (PPCs) and human experts. Our main findings indicate that: 1) users' trust in human-generated and machine-generated ratings of password strength is heavily influenced by users' own subjective judgments; 2) users behave differently for different passwords; 3) there are different behavioral patterns that can strongly influence users' decisions; 4) users have a (slightly) higher tendency to trust machine-generated ratings when their own subjective judgments match the machine-generated ratings. We hope that this reported work can stimulate more research into this less investigated area of password research.

Acknowledgement. The authors would like to acknowledge and highlight that Nouf Aljaffan, a former PhD student at the University of Surrey in the UK, was the actual lead contributor of the reported work, as part of her PhD study. She could not be listed as a co-author because the other co-authors lost contact with her since she left the University of Surrey. Rather than listing her as a co-author without her explicit consent, the other co-authors felt it is more appropriate to credit her contribution in this section. The work reported in this paper appeared as Chapter 3 of her thesis [41], so her contribution is clearly evidenced there. Furthermore, this work was funded by Nouf Aljaffan's PhD scholarship from the King Saud University, Saudi Arabia.

The authors would also like to thank Sebastian E. Bartos, who was working at the University of Surrey during this research, for his help for the statistical analysis reported in this paper.

The work of the last three co-authors was partly supported by the UK part of a joint Singapore-UK research project "COMMANDO-HUMANS: COMputational modeling and Automatic Non-intrusive Detection Of HUMan behAviour based iNSecurity" (http://www.commando-humans.net/), funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/N020111/1.

References

[1] Bonneau, J. (2012) The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *Proc. S&P 2012*: 538–552.

[2] Bonneau, J., Herley, C., van Oorschot, P.C. and Stajano, F. (2015) Passwords and the evolution of imperfect authentication. *Communications of the ACM* **58**(7): 78–87.

[3] Florêncio, D. and Herley, C. (2007) A large-scale study of web password habits. In *Proc. WWW 2007*: 657–665.

[4] Li, Y., Wang, H. and Sun, K. (2016) A study of personal information in human-chosen passwords and its security implications. In *Proc. IEEE INFOCOM 2016*: 1–9.

[5] Carnavalet, X.D.C.D. and Mannan, M. (2015) A large-scale evaluation of high-impact password strength meters. *ACM Transactions on Information and System Security* **18**(1): 1:1–1:32.

[6] Ur, B., Kelley, P.G., Komanduri, S., Lee, J., Maass, M., Mazurek, M.L., Passaro, T. *et al.* (2012) How does your password measure up? the effect of strength meters on password creation. In *Proc. USENIX Security 2012*: 65–80.

[7] Egelman, S., Sotirakopoulos, A., Muslukhov, I., Beznosov, K. and Herley, C. (2013) Does my password go up to eleven? the impact of password meters on password selection. In *Proc. CHI 2013*: 2379–2388.

[8] Dropbox, Inc. (2015), zxcvbn: A realistic password strength estimator, https://github.com/dropbox/zxcvbn/.

[9] Burr, W.E., Dodson, D.F., Newton, E.M., Perlner, R.A., Polk, W.T., Gupta, S. and Nabbus, E.A. (2013), Electronic authentication guideline, NIST Special Publication 800-63-2.

[10] Wang, D., He, D., Cheng, H. and Wang, P. (2016) fuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars. In *Proc. DSN 2016*: 595–606.

[11] Narayanan, A. and Shmatikov, V. (2005) Fast dictionary attacks on passwords using time-space tradeoff. In *Proc. CCS 2005*: 364–372.

[12] Castelluccia, C., Dürmuth, M. and Perito, D. (2012) Adaptive password-strength meters from Markov models. In *Proc. NDSS 2012*.

[13] Weir, M., Aggarwal, S., de Medeiros, B. and Glodek, B. (2009) Password cracking using probabilistic context-free grammars. In *Proc. IEEE S&P 2009*: 391–405.

[14] Wang, D., Zhang, Z., Wang, P., Yan, J. and Huang, X. (2016) Targeted online password guessing: An underestimated threat. In *Proc. CCS 2016*: 1242–1254.

[15] Melicher, W., Ur, B., Segreti, S.M., Komanduri, S., Bauer, L., Christin, N. and Cranor, L.F. (2016) Fast, lean and accurate: Modeling password guessability using neural networks. In *Proc. USENIX Security 2016*: 175–191.

[16] Javier Galbally, I.C. and Sanchez, I. (2017) A new multimodal approach for password strength estimation. Part I: Theory and algorithms. *IEEE Transactions on Information Forensics and Security* **12**(12): 2829–2844.

[17] Ur, B., Alfieri, F., Aung, M., Bauer, L., Christin, N., Colnago, J., Cranor, L.F. *et al.* (2017) Design and evaluation of a data-driven password meter. In *Proc. CHI 2017*: 3775–3786.

[18] Sotirakopoulos, A., Muslukov, I., Beznosov, K., Herley, C. and Egelman, S. (2011) Poster: Motivating users to choose better passwords through peer pressure. In *Proc. SOUPS 2011*.

[19] Sotirakopoulos, A. (2011) *Influencing User Password Choice Through Peer Pressure*. Master's thesis, University of British Columbia, Canada.

[20] Aljaffan, N., Yuan, H. and Li, S. (2017) PSV (Password Security Visualizer): From password checking to user education. In *Proc. HAS 2017 (HCII 2017)*: 191–211.

[21] Cheng, X., Fu, S. and de Vreede, G.J. (2017) Understanding trust influencing factors in social media communication: A qualitative study. *International Journal of Information Management* **37**(2): 25–35.

[22] Mayer, R.C., Davis, J.H. and Schoorman, F.D. (1995) An integrative model of organizational trust. *Academy of Management Review* **20**(3): 709–734.

[23] Seitz, T. and Hussmann, H. (2017) PASDJO: Quantifying password strength perceptions with an online game. In *Proc. OzCHI 2017*: 117–125.

[24] Huang, D.L., Rau, P.L.P., Salvendy, G., Gao, F. and Zhou, J. (2011) Factors affecting perception of information security and their impacts on IT adoption and security practices. *International Journal of Human-Computer Studies* **69**(12): 870–883.

[25] Costante, E., den Hartog, J. and Petkovic, M. (2011) On-line trust perception: What really matters. In *Proc. STAST 2011*: 52–59.

[26] Willis, J. and Todorov, A. (2006) First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science* **17**(7): 592–598.

[27] Brambilla, M., Rusconi, P., Sacchi, S. and Cherubini, P. (2011) Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology* **41**(2): 135–143.

[28] Brambilla, M. and Leach, C.W. (2014) On the importance of being moral: the distinctive role of morality in social judgement. *Social Cognition* **32**(4): 397–408.

[29] Goodwin, G.P., Piazza, J. and Rozin, P. (2014) Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology* **106**(1): 148–168.

[30] Westerman, D., Spence, P.R. and Van Der Heide, B. (2014) Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication* **19**(2): 171–183.

[31] Toma, C.L. (2014) Counting on friends: Cues to perceived trustworthiness in Facebook profiles. In *Proc. ICWSM 2014*: 495–504.

[32] Ur, B., Noma, F., Bees, J., Segreti, S.M., Shay, R., Bauer, L., Christin, N. *et al.* (2015) "I added '!' at the end to make it secure": Observing password creation in the lab. In *Proc. SOUPS 2015*: 123–140.

[33] Ur, B., Bees, J., Segreti, S.M., Bauer, L., Christin, N., Cranor, L.F. and Deepak, A. (2016) Do users' perceptions of password security match reality? In *Proc. CHI 2016*: 3748–3760.

[34] Lynn, L.A. (1978) *Language Emotionality, Source Credibility, and Sex Effects: An Experimental Study of Communication Perception*. Phd thesis, Department of Speech, Indiana University, USA.

[35] William K. Darley, R.E.S. (1993) Advertising claim objectivity: Antecedents and effects. *Journal of Marketing* **57**(4): 100–113.

[36] Dietvorst, B.J., Simmons, J.P. and Massey, C. (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* **64**(3): 1155–1170.

[37] Chen, J., Gates, C.S., Li, N. and Proctor, R.W. (2015) Influence of risk/safety information framing on Android app-installation decisions. *Journal of Cognitive Engineering and Decision Making* **9**(2): 149–168.

[38] Chong, I., Ge, H., Li, N. and Proctor, R.W. (2018) Influence of privacy priming and security framing on mobile app selection. *Computers & Security* **78**: 143–154.

[39] Sundar, S.S. (2008) The MAIN model: A heuristic approach to understanding technology effects on credibility. In *Digital Media, Youth, and Credibility* (The MIT Press), 73–100.

[40] Wheeler, D.L. (2016) zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security 2016*: 157–173.

[41] Aljaffan, N.M.D. (2017) *Password Security and Usability: From Password Checkers To a New Framework For User Authentication*. Phd thesis, Department of Computer Science, University of Surrey, UK.

Appendix A. Predefined Reasons

Here, we list the list of predefined reasons which depend on the user's choice of rating.

If a user selected ***"objective rating"***, he or she saw the following options of reasons.

1. Software can detect hidden things which users cannot.
2. Users often tend to make mistakes while software is more accurate.
3. I do not trust subjective rating as I think not all users have a good experience in the field.
4. I selected that rating because it matches my expectation.
5. I selected the lower rating to be safe.
6. I used my own experience/knowledge to make a judgment.
7. Others.

If a user selected ***"subjective rating"***, he or she saw the following options of reasons.

1. I trust users because software cannot predict new form of attacks.
2. Applications are created by human, so it is better to trust user rating.
3. I think software can produce a misleading rating since software might be compromised or not designed well.
4. I selected that rating because it matches my expectation.
5. I selected the lower rating to be safe.
6. I used my own experience/knowledge to make a judgment.
7. Others.

If a user selected ***"neither"***, he or she saw the following options of reasons.

1. None of the two ratings match my expectation.
2. I consider both options in addition to my experience to form my own rating.
3. Others.

If a user selected ***"undecided"***, he or she saw the following options of reasons.

1. I need more details to make a proper decision.
2. Others.