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}.

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

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".

[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

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".

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)).

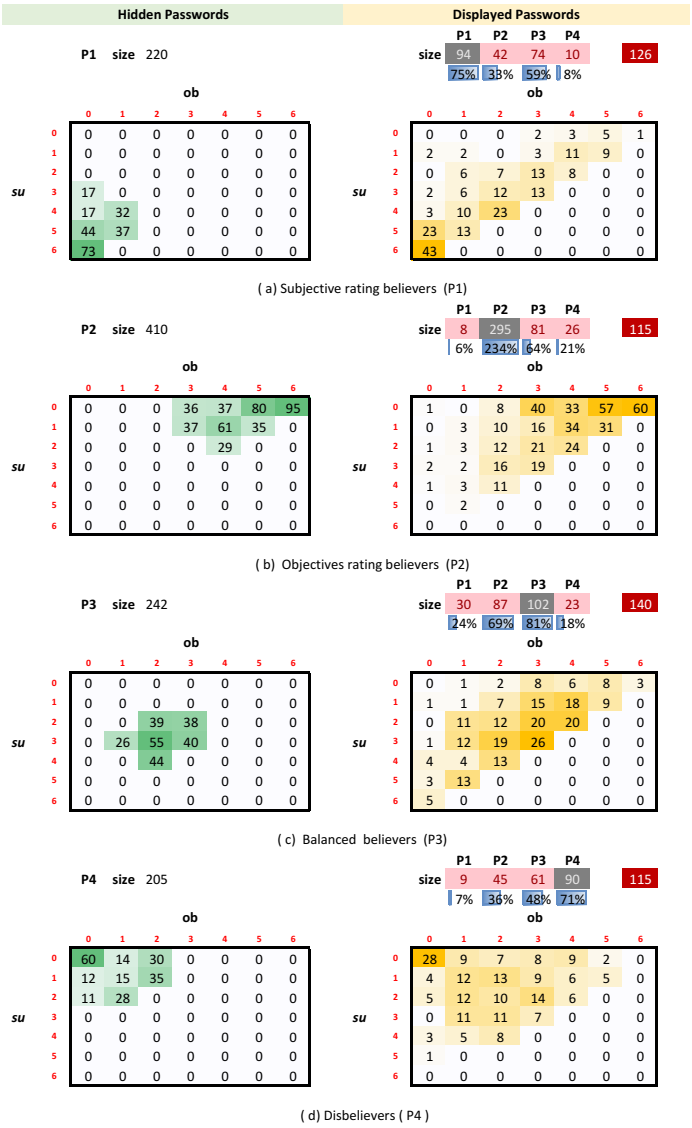


Figure 4. 2-D distribution of participants' responses according to behavioral patterns and password display conditions.

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.

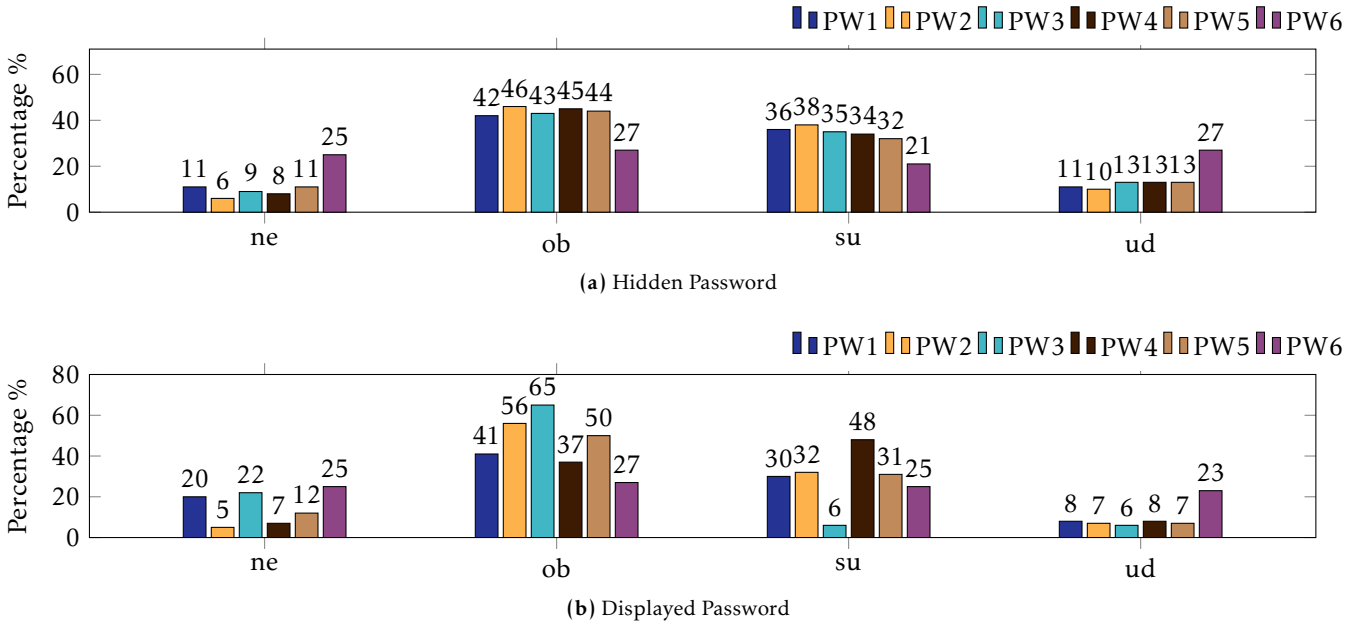
The above results can also be seen visually in Figure 5, which shows percentages of participants' answers for each of the four options and for all the five passwords. Figure 5(a) shows that participants had a convergence of views in terms of trust when passwords were hidden. However, this is not the 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.