Intelligo ut Confido: understanding, trust and user experience in verifiable receipt-free e-voting

Marie-Laure Zollinger[1], Peter B. Rønne[1], Steve Schneider[2],
Peter Y. A. Ryan[1] and Wojciech Jamroga[3]

[1] Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
[2] Surrey Centre for Cyber Security, University of Surrey, UK
[3] Institute of Computer Science, Polish Academy of Sciences

Abstract. Voting protocols seek to provide integrity and vote privacy in elections. To achieve integrity, procedures have been proposed allowing voters to verify that their vote is correctly counted– however this impacts both the user experience and privacy. In particular, vote verification can lead to vote-buying or coercion, if an attacker can obtain a proof of the cast vote. Thus, some voting protocols provide mechanisms to prevent such receipts. To be effective, such *receipt-freeness* depends on voters being able to understand and use these mechanisms. In this paper, we present a study with 300 participants to evaluate the voters' experience and understanding of the receipt-freeness procedures in the Selene scheme in the context of vote-buying. This is the first user study dealing with vote-buying in e-voting. While the usability and trust factors were rated low in the experiments, we found a positive correlation between trust and understanding.

1 Introduction

Voting and elections are a prime example of socio-techncial systemswhere humans interact in a technological environment [6]. This applies even more obviously to electronic voting [11]. *Voting protocols* are designed to satisfy certain important properties, in particular Privacy and Integrity. Privacy is often defined by three sub-properties: Ballot-Secrecy, Receipt-Freeness and Coercion-Resistance. Ballot-Secrecy ensures that the protocol does not reveal the voter's choice. Receipt-Freeness says that the system will not provide any evidence enabling a voter to prove how they voted. Finally, Coercion-Resistanceanables the voter to pretend to cooperate actively with a coercer [8], but still cast their intended vote. When interacting with a vote-buyer, a voter has an economical incentive to obtain a receipt of the vote. A vote buyer offers a voter money for a vote cast a particular vote, but the money is only paid upon receiving "proof" of the vote. However, if the "proof" can be faked, the vote buyer cannot trust the receipt and hence vote buying should be disincentivised.

Integrity means that the announced outcome of the election is correct. Verifiable schemes demand more: the system should also deliver a proof that the result is correct. *End-to-end verifiable voting protocols* [28], entail two complementary procedures: firstly, universal verifiability means that anyone can check that the vote count is correctly computed from the submitted ballots, secondly, individual verifiability means that each voter can check that their vote intent was correctly captured in the submitted ballot. The latter is most interesting from a user perspective since it inherently involves user interaction.

In the Selene e-voting protocol [27] voters receive a tracking number which points to their vote in plaintext in the tally. Voters can present a fake tracking number to a vote-buyer, providing a *receipt-free mechanism.* A vote buyer cannot determine whether the presented tracker is real or fake, and hence has no proof of how the voter actually voted. The voter's understanding and the user experience of the verifiability procedures in Selene were explored in several papers [9, 34, 22, 35]. However those studies did not include the receipt-free mechanism which introduce additional trust issues.

Receipt-Free or Coercion-Resistance mechanisms have rarely been tested with end-users; to our knowledge, only [23] explored a Coercion-Resistance mechanism for the JCJ e-voting protocol [12], and Receipt-Freeness in the context of Vote-Buying has not been investigated. This is a gap in the assessment of practical security of voting procedures. For an overview, see [14].

In this paper, we present the first large scale study of the receipt-free mechanism iof the Selene voting protocol. The study is based on experiments with 300 human participants recruited through the platform Prolific. We evaluated the user experience (UX), trust, and understanding of the voting procedure, and formulated three hypotheses to be tested:

H1 The voting application and its receipt-free feature provide a positive user experience to the participants.
H2 The application and receipt-free mechanism are trusted by the participants.
H3 Participants who understand the receipt-free mechanism have increased trust in the application.

To evaluate the UX, we use the user experience questionnaire (UEQ). At the time of the user experiment there was no standard questionnaire to assess this metric in the voting context (Ref. [1] appeared later). Therefore, we defined trust for voting and proposed a new questionnaire assessing the voters' trust in the protocol, see Section 4. Correct understanding of the receipt-free mechanism was evaluated by observing the steps performed by participants. To evaluate understanding, we designed game inspired by [18] for privacy in voting. Correct understanding of the mechanism leads to a specific workflow, see Section 5.

Finally, participants were invited to tell us why they made their choice in the game, and how they felt. We categorized their answers in a qualitative analysis and correlated this with the participants' understanding (Sec. 6.2).

To summarize, our contributions are:

- A questionnaire to evaluate trust in the context of voting,
- A unique game design to assess the voters' understanding of a system,
- An evaluation of the relationship between understanding and trust,

- A qualitative analysis of user feedback on receipt-freeness and vote-buying,
- A list of recommendations for future voting systems and user studies.

2 Related Work

Our experiment is inspired by [18] where a game approach was used to evaluate the understanding of the privacy mechanisms in the e-voting protocol Prêt-à-Voter (PaV) [26]. In PaV voters get receipts, but with their votes in encrypted form. In the game the 12 participants tried to guess each other's votes and had the choice between publishing their receipt or not. They were rewarded for revealing it. Hence, participants who understood that receipts did not reveal their vote should choose to reveal the receipt as the most profitable strategy. Thus understanding could be measured, but with so few participants a conclusion was hard to draw. We improve on this with a large number of participants.

Until now, most studies focused on the usability and appreciation of voters for a given system, but an evaluation of their understanding is rarely performed. Also, it has been been evaluated with reference to predefined mental models of the participants. In [2], the authors let voters draw their mental models for three voting schemes. This study reveals that voters focused much more on the voting phase in all three protocols, as the verification features remained unclear to them. In the case of Selene, two studies have looked at mental models of participants [34, 35]. It appears that the understanding of verification was better when the participants have seen a possible threat, e.g. a vote manipulation [35]. The verification mechanisms of Selene were implemented without the receipt-free mechanism [29], augmenting an existing voting system. The user experience was evaluated [4] showing satisfaction and a higher confidence in the system. The evaluation of coercion-mitigation features have rarely been performed, except for the protocol JCJ [12] in [23].

3 The Selene protocol

Selene is an e-voting protocol designed to make the individual verification more usable and intuitive for voters. Verification procedures can be categorized into four types, [22]: audit-or-cast, verification device, code sheets and tracker-based. Selene belongs to the last category, the other categories require the voters to either handle ciphertexts, or to verify codes.Tracker based protocols allow voters to verify the presence of their vote in plaintext in the final tally using a (private, deniable) tracking number. The special feature of Selene is that this tracker is only delivered to the voter *after* the tally is published to allow the coercion evasion strategy described below.

The complete description of the protocol is available at [27]. Each voter has a pair of public and private keys that are used in the verification phase. The election keys are also generated and the election public key is distributed to voters. A public bulletin board (BB) is used to display the public data.

1) Setup. The election authorities generate the list of tracking numbers. These are encrypted under the election public key, then shuffled and associated with the voters. A trapdoor commitment to each tracker is created and published on the bulletin board, sealing the relation between a tracker and a voter. To open a commitment and see the tracker, one needs the voter's private key and a secret (dual key) which is revealed later to the voter by the authorities.

2) Voting. When the setup phase is over, voters can cast a vote encrypted with the election public key. The encrypted vote is published on BB.

3) Tally. After voting, the authorities retrieve the pairs of encrypted tracking numbers and encrypted votes, shuffle the pairs and decrypt them to obtain and publish the pairs of plaintext tracking numbers and votes.

4a) Verifying. Some time after the tally is published, the secret dual key associated to each commitment is delivered to the voter. Combining the dual key, the commitment and their private key, each voter can retrieve the tracking number, and verify the associated plaintext vote.

4b) Faking. If a voter is interacting with a vote-buyer or being coerced, the voter can choose an alternative tracker, showing a plaintext vote that corresponds to the adversary's request. From this tracker and the commitment, a fake dual key is computed by the voter using her private key. This can be done after the tally phase. The combination of this fake dual key with the commitment and private key of the voter will open to the selected fake tracking number.

In the trial, participants could verify their own vote and later request that an alternative tracker be displayed to mislead the vote-buyer. This results in a more complex experience compared to what most voters would encounter in normal elections.

Web application For the experiment, we implemented a web app reflecting the user steps described above. The voter can access the following pages through a menu, after login:

- *Home:* this page explains the purpose of the web app and the different pages.
- *Voting:* the voting question is displayed with the possible vote choices.
- *Verification:* this page presents the election result as vote/tracker pairs. The voter can retrieve the tracking number to verify the vote, or choose a fake tracking number.
- *About:* information about Selene and its features is displayed here.
- *Contact:* a link to our email is provided in case of questions.
- *Logout:* used to log out from the study.

A default workflow is proposed once the voter is connected. In the voting section, after selecting the candidate, a confirmation page is displayed. The voter can the click on a button "Encrypt and send my vote". As shown in [35, 22], such an interaction does not require any skill, but increases the security perception. On the verification page, the tally is displayed and the voter is offered two choices: fake the tracking number in case of coercion or vote-buying, or go for verification directly. To fake the tracking number, a new page is displayed where the voter can access the bulletin board and type the chosen tracking number. The voter is warned that it is not possible to retrieve the real tracking number after this request. After validating, the voter is redirected to the main verification page. If the voter chooses to verify, the app computes the (real or fake) tracking number and the voter can connect to the bulletin board to verify the vote.

In previous implementations of Selene [34, 35, 9, 22], the authors decided to highlight the tracking number and corresponding vote directly in the application to increase usability, with the risk of lowering privacy and the security perception. In this version, we provide the tracking number and the user has to display the bulletin board and look for the tracker to verify the vote. This is more faithful to the original protocol design but less usable.

4 Trust

Trust features in many studies about voting [7, 33, 30, 15, 21, 34, 3]. It is rather complex to evaluate, as trust has many aspects: trust in politics, trust in digital technologies, understanding of the app, etc.

There is no standard questionnaire available to evaluate trust of users for voting systems. The UEQ+ questionnaire [32] proposes little related to trust. To close this gap and explore the relation between understanding and trust, we designed a more specific questionnaire for the e-voting context. We now discuss trust and the design of the questionnaire. After our experiment was done another trust measure for voting was proposed [1]. However, with 44 questions this is not suitable for our online experiment where participants have limited patience.

In [19], Luhmann differentiates trust and confidence. Confidence can be obtained without any additional explanation, in particular security does not need to be perceived to be acknowledged while trust requires an evaluation from the users' of their security perception to be granted.

In [24], Pieters observes that a voting system can obtain the voters' *confidence* if it works correctly. A system that guarantees a correct result should not worry the voters. But, when a new system implementing new procedures, such as verifiability features, is comparised to the old system which has the confidence of voters, *trust* may be impacted. The author also mentions the relationship between trust and explanation. The voters need to *understand* verifiability in a new system to convince them to use it. Previous works have already mentioned the relationship between trust and the explanations [10], and in voting [34, 35].

We aimed to provide a reasonable amount of information about the protocol, to support a good trust rating. However, the participants have limited time to evaluate the app, so we should not provide too much information that could overwhelm them.

4.1 Our metric

Our voting-oriented trust questionnaire contains eight questions. From the studies and literature cited above, we see that trust depends on a positive evaluation of the security. In our questionnaire, we evaluate the feeling of security on one hand; and the acceptance of the system on the other, to see if trust is engendered. The questions, labelled by topic, are 1) [Acceptance] “I trust the system and I would use it in a real election”. 2) [Security] “I believe that the personal information (vote included) is kept private”. 3) [Security] “I think that the system ensures the integrity of the elections”. 4) [Security] “I think that the system is transparent and lets me know everything about its behaviour”. 5) [Acceptance] “I think that the verification phase is important”. 6) [Security] “I was convinced by the verification phase that my vote was correctly recorded”. 7) [Acceptance] “I would use such a verification system if it was available”. 8) [Security] “I think that the result of the election can be changed by an attacker”. Answers were given on a Likert scale with 6 choices from strongly disagree to strongly agree. The results were scaled so that each question gives 0-10 points, with 10 indicating maximal trust. We used the following classification: High trust for a score > 64, moderate trust $48 - 63$, low trust $32 - 47$ and very low trust < 32.

5 User protocol

For the experiment we used the crowd-sourcing platform Prolific [25]. The context provided to the participants was the following: the city council is organising local elections to request its citizens’ opinion on several society subjects. To cast their vote, the participants used our online application.

Trust and UX were evaluated in the standard way: after having interacted with the application, the participants were given questionnaires. We used the System Usability Scale, the User Experience questionnaire [31] and our trust questionnaire. Then, to evaluate the understanding, we designed a user game inspired by the game theoretic experiment in [18]. The participants interact once more with the application but we provided an additional scenario: the participant had to interact with a vote buyer[4]. The instructions from the vote buyer were displayed in a box next to the web page: the vote-buyer asks for a different vote than the choice made by the voter (we configured the game by asking in advance the voter’s opinion, see below). Our evaluation consists in looking at the participant’s behaviour in such a scenario. Our assumption is that a correct understanding will lead the participants to vote for their candidate and use the receipt-free mechanism to provide a fake tracker to the vote buyer.

Pilot studies. We ran two pilot studies with five participants in each. In the first pilot, none of the five participants watched the video nor tried the receipt-free mechanism (even with the vote-buying scenario) and they finished the study in less than five minutes (while 20 min. were given). This rush bias is well known and called “satisficing” in Prolific’s terms of use. To ensure the participants use the app fully, we introduced a workflow: they could not access the questionnaires and continue the study before they used the mechanism to get a new tracking number. Guidance was provided as side notes on the website. Also, some attention checks

[4] With Selene, countering vote buying and coercion involves the same user steps.

were added to the questionnaires as recommended by Prolific. We further discuss the limitations in section 5.1 below.

Participants were paid 2.5£ for the study (20 minutes) which was evaluated as a *Good* hourly rate by Prolific, and we added an extra 1£ as a bonus payment for having played the game.

After a consent form, the user experiment had the following steps

Demographics We recruited 300 participants on the crowd-sourcing website Prolific [25]. We used the pre-screening feature to select participants: to ensure that they have a similar experience in voting, we chose UK citizens living in UK. The average age was 33 years (Min=18, Max=73, SD=11). They come from various backgrounds, the education level differed: No diploma (0,67%), A-Levels (13,33%), College Level (19,33%), Bachelor (42,33%), Master Degree (20%), PhD (1,33%) and other (3%). Finally, regarding their attitude toward online voting, 2,33% were negative, 7% were rather negative, 39,67% were neutral, 35,67% were rather positive and 14,67% were positive.

Configuration In the end of the demographics' questionnaire, we asked the participants to answer the voting question used in the game, to configure the vote buyer's instruction. The question was about the COVID-19 crisis:

Regarding the recent events related to the COVID-19 pandemic, according to you, what would be the best policy to adopt at the beginning of the epidemic?
- A strict confinement for all
- No confinement but detection tests available for everyone

We configured the game by changing the vote buyer's instructions according to their opinion. If they chose "A strict opinion for all", the vote buyer asks for "No confinement but detection tests available for everyone" and vice versa.

Video explaining the protocol: We describe the Selene protocol in a 4-minute video that the voter was invited to watch.

A tutorial to demonstrate the receipt-free mechanism First, we let the participants use the application through a tutorial. As mentioned above, the first pilot study has shown that participants were rushing to end the study as fast as possible. The tutorial ensures that they see and test all available features in the application, a specific workflow was forced with guidance, given as side notes. Therefore, participants were able to verify their vote and then fake their tracking number. We wanted to show that they can see their plaintext vote, but also have the ability to change their tracking number to show another vote to a coercer or vote-buyer.

Questionnaires We evaluated the usability, user experience and trust after this tutorial phase. The reason was that we did not want to influence their trust rating by going through a coercive scenario, but obtain their general impression of the app. Also, we put a few attention checks (through questions about the app) at the beginning of our questionnaire. The checks were announced in the study description on Prolific. Our goal was to increase the attention given to the explanations in the app. Of course there is a possibility that the participant did not understand the protocol and provide wrong answers. We did not exclude such participants, our goal was to help them to focus on the information rather than skipping it as in the pilot study.

Vote-Buying Game We introduced the game by telling the participants that they will receive instructions from a vote buyer. The rules were given as:

A vote buyer wants to buy your vote by giving you a vote instruction. He may ask you how you voted and to reveal your tracking code, in which case you can give an alternate code.
If you send a tracking code for the requested candidate, you will receive 70 pence from the vote buyer.
If you want to keep your vote intention, you will receive 30 pence.
These incentives will be provided as bonus payment after the study.[5]

When participants start, they were asked to vote as they did in the tutorial but additional instructions given by the vote buyer on the left side of the screen. The participants van choose whether or not to follow the vote buyer's instructions. Our idea was to determine whether the participants understood that they can keep their vote while convincing the vote buyer that they follow his choice. Indeed, the dominant strategy for a player, given the possibilities offered by the application, is to cast the intended vote while selling a fake tracker to the vote buyer.[6] After computing the tracking number, the participant could choose to send it to the vote buyer or not by clicking on a button.

End of Study To finish the study, the participants were asked to tell which choice they made - keep their vote intention or follow the vote buyer's instructions - and why. Our last question was about how they felt during the game.

Ethical approval We obtained ethical approval from our institution's Ethics Panel. Our work is compliant with GDPR and the research terms of Prolific.

5.1 Limitations

While Prolific brought many advantages, including reaching many participants rapidly, and good demographic samples, we found some limitations.

Regarding our trust questionnaire, even though we built the questionnaire to answer specific needs, we are aware that the questionnaire needs further testing to be validated by the community. This first study using it is an attempt to grasp insights on trust with a specific approach of security perception and acceptance.

Correlations were shown between our measurements (see the next section): some items have been assessed *before* the vote-buying game (trust, usability), while others have been asked *after* the vote buying game (feelings). The correlations found between those measurements could be altered by the game.

[5] In the end we provided both incentives as bonus payment to all participants regardless of their choice, for fairness.

[6] Note that the instructions were formulated without directly revealing this optimal strategy, but the participants should deduce it if they understood the introduction to the study and the explanatory video.

Our first pilot study showed that participants are rushing, likely to increase their reward per hour. Without any guidance, we could not hope that participants will visit all pages in our app, forcing us to make them first test the app through a tutorial rather than exploration. This is known as “satisficing bias” and is acknowledged by Prolific [25]. To counter this, we asked the participants to answer questions regarding their understanding in the app: these “attention checks” are recommended by Prolific and helped us to lower this bias.

Another limitation concerns our scenario with vote-buying. As for studies in the lab, participants might have a bias to give a good image of themselves, hence answering what would be ethically acceptable [16, 17]. In this study, some participants justified themselves for having followed the vote buyer because “this is just a game”, or mentioned their integrity for not having followed him.

Finally, we ask participants to understand new features in a limited amount of time. More time would be necessary to understand the features.

6 Results: Evaluation of Understanding of receipt-freeness

6.1 Quantitative results

Usability and User Experience In this section we will explore the results obtained for the user experience and the usability questionnaires. Following to the UX handbook [31], a result above 0.8 for the UEQ categories would be considered as positive.

We obtained the following results with the UEQ: Attractiveness obtained -0.1 (SD=0.08), Perspecuity obtained -0.41 (SD=0.09), Efficiency obtained 0.31 (SD=0.09), Dependability obtained 0.6 (SD=0.06), Stimulation obtained 0.12 (SD=0.07), Novelty obtained 0.55 (SD=0.07).

Compared to the previous studies on Selene measuring the user experience through a mobile application [9, 22], we can see that the web application performed poorly. The attractiveness has been rated as -0.1 (SD=0.08), the usability aspects received the score of 0.16 (SD=0.08) and the hedonic aspects received the score of 0.33 (SD=0.06). At a subscale level, dependability received the higher score with 0.6 (SD=0.06).

Where perspicuity (difficult to learn/easy to learn) was the highest score in [9] (with 2.16 and 1.90), we obtained the lower score with -0.41 (SD=0.09). We will discuss the possible reasons in the discussion.

We summarise the SUS (System Usability Scale) results. We measured effectiveness by asking the participants to give a self assessment of their individual verification step: we asked if they found their tracking code on the bulletin board. Only 86% of the participants answered that they found their vote, even though we know that all participants have computed their tracking number.

Efficiency was measured through the time taken by the participants to vote and to compute their tracking code after having logged in to the application. The mean time is 57 seconds (median=45.5, SD=39.65, min=17, max=324).

Compared to [22], again, the web application performed poorly on the satisfaction scale (mean=48.67, median=45, SD=22.81, min=0, max=100) with a mean score below 51, considered as “unacceptable” in [5]. We can also note that participants were on average six times faster to vote and verify compared to the lab study in [22], while the minimum time to cast a vote is almost twelve times faster with the web app, questioning the participants’ commitment to the test.

In conclusion, the hypothesis **H1** is not supported by the experiments: our web app did not provide a positive user experience (scores below 0.8) nor an acceptable usability.

Trust As mentioned above, the questionnaires were filled after the tutorial phase and before the game. This was to let the participants give an evaluation of the app and of its features before we collect the data regarding their understanding. We did not want a specific threat scenario to influence their opinion on the protocol itself.

Overall, trust received an evaluation of 46.81 ($SD = 16.132, Min = 4, Max = 78$). On the subscale level, the acceptance (over 30) was rated 18.59 ($SD = 7.264, Min = 0, Max = 30$) and the feeling of security (over 50) was rated 28 ($SD = 20.093, Min = 0, Max = 48$).

Regarding the grading proposed in section 4, the trust has been evaluated as low by the participants. We can conclude from this result that our hypothesis **H2** is not supported by our results.

Understanding As a reminder, we evaluate the understanding of the receipt-free mechanism as correct when participants kept their vote intention while faking their tracker for the vote-buyer. In total, 54 of 300 participants chose this dominant strategy, and *correctly* understood the faking mechanism.

6.2 Qualitative results and relations between variables

We have done a qualitative analysis of the answers from the game and the feedback from our two last questions. The details are included in the full version of the paper[7]. Especially we categorise the answer to the question “Why have you made this choice in the game?” in terms of the labels money, integrity, understanding, experimenting (wanting to experiment) and miscellaneous. And for the question “How did you feel during the study” we use the labels overwhelmed, stressed, offended, good, interested, confident, confused and observed.

While the questionnaires were filled after the first phase (tutorial) of the user study, the understanding of participants and the qualitative data were collected after the second phase (game). In particular, the vote-buying scenario might have impacted some participants’ feedback especially their feeling regarding the study. The following correlations should be considered under this limitation.

Trust and Understanding: When defining trust, our questionnaire was built with the idea that the explanations provided were important to give transparency and to increase the voters’ understanding in the application. During the study, we gave explanations through video and text, participants followed a tutorial before playing a game designed to evaluate their understanding of the

[7] https://arxiv.org/abs/2407.13240

features. This study design allows us to check the correlation between the Trust results and the voters' Understanding, measured by observing their decisions.

Understanding was measured by looking at the capacity of a participant to vote as intended while faking the tracker for the vote buyer. We obtained one group of 54 participants out of 300 who understood. To measure the correlation between trust and understanding, we performed an independent t-test. The participants who understood the concealing feature gave a statistically higher evaluation of trust ($Mean = 51.22$, $SD = 15.372$) compared to participants who did not understand it ($Mean = 45.84$, $SD = 16.163$), $t(298) = 2.236$, $p = 0.026$. Further, Cohen's effect size value ($d = 0.34$) suggested a small to moderate practical significance. We conclude the evidence was in favour of hypothesis **H3**.

Trust and Satisfaction Measures: We computed the Pearson correlation coefficient $r = 0.561$ ($p = 0.01$) between our trust and satisfaction measures, implying a moderate positive correlation between trust and usability. Similarly, the coefficients between Trust and the UEQ's scale are: Attractiveness 0.14 (p=0.05); Perspicuity 0.135 (p=0.05); Efficiency 0.149 (p=0.01); Dependability 0.151 (p=0.01); Stimulation 0.173 (p=0.01); Novelty 0.063. The values for r are below 0.2 indicating a weak positive relation.

Understanding and Time Spent in the Study: 47 participants finished in less than 20 min (which was the planned time), whereas the mean was 35 min and 55 sec. Participants took more time than planned, probably because of our attention checks, added after the pilot studies where participants rushed through within five minutes. We ran a one-way ANOVA test, which showed no significant difference between those who understood the game and the others.

Self-explanation/Feeling and Understanding: Of the 54 participants who faked their tracking code to send to the vote buyer, 26 mentioned integrity, 3 money, 17 gave an explanation about their understanding. Conversely, 2 participants explained correctly how the system works, but did not fake their tracking code for the vote buyer. Regarding feelings, 22 participants of the 54 said that they were confused, 25 that they were felt good, confident or interested in the system, the remaining 7 were felt observed, stressed, overwhelmed or frustrated.

A Welch ANOVA test between the decision categorization and the understanding shows no significant differences between the five groups ($p > 0.05$). Hence in our sample, we cannot conclude on the relation of the understanding of participants to the reason for following the vote buyer or not.Similarly, we found no significant differences between the 8 groups of feelings ($p > 0.05$). Thus the understanding of participants might not be related to the feelings of participants.

Self-explanation/Feeling and Trust: The relation between the decision's categories and the trust assessments is analyzed with a 1-way ANOVA. The ANOVA test shows a significant difference between the five categories ($F(4, 295) = 2.872, p = 0.023$). A post-hoc Tukey is run to locate differences between categories, and found that participants who mentioned integrity rated trust better (8 points) than those interested in money ($p = 0.016$). On the other hand, there was no significant difference between the 8 groups of feelings ($p > 0.05$). Hence, the participants' trust (evaluated after the tutorial) was not influenced by their feelings (evaluated after the game).

Self-explanation/Feeling and Usability: We run a 1-way ANOVA test to investigate a relation between the SUS assessments and the self-explanation provided. The test shows a significant difference between the five categories ($F(4, 295) = 2, 729, p = 0.029$). A post-hoc Tukey found that participants who mentioned an experimentation gave a better evaluation than those doing the test for money ($p = 0.049$).

We also run a 1-way ANOVA test to find a relation between the feeling's categories and the SUS assessments. The ANOVA test shows a significant difference between the 8 categories ($F(7, 292) = 3.446, p = 0.001$). A post-hoc Tukey found that participants who felt interested rated better than those feeling overwhelmed or stressed (for reference here $p < 0.05$). The details of the analysis are as follows (we report those with a significant difference only):

	Difference between the means	P value
Experimenting over Money	15.48	0.049
Interested over Overwhelmed	21.34	0.039
Interested over Stressed	21.34	0.009

Similarly, we run a 1-way ANOVA to find relations between the UEQ items and the categories for self-explanation and feelings. For self-explanation, no relation was found ($p > 0.05$). We found a relation between the feelings' groups and the UEQ items with statistical significance ($p < 0.001$). Overall, participants having a positive feeling regarding the app rated it better than the other participants with $p < 0.05$.

6.3 Analysis and Discussion

No relation was found between the understanding of participants and their self-explanation or feeling regarding the application. However, we have seen that trust and understanding are correlated, which supports our hypothesis.

We see that the user experience and usability were poorly rated. Here we found a moderate correlation between satisfaction and trust, but only a small correlation between UEQ items and trust. In the SUS questionnaire, some items concern the acceptance of the tested application, which is one aspect of our trust questionnaire, and might explain the stronger correlation. However, we argue that a good user interface will benefit a voting application. In [20] and [13], the authors mention the signals impacting trust, including usability. We had good results regarding effectiveness and the efficiency, but we failed at convincing the participants that our application was easy to use and enjoyable.

To explain this, we look at the feelings formulated by the participants. The most expressed feeling was *confusion*: with participants unsure about the steps to follow. A highlighted reason was the complexity of the study, while Prolific's users are used to surveys, which are linear and require less commitment (in the sense of direct interactions influencing the behaviour of the app) from the user. Other feelings expressed by participants were *stress* and *frustration*.

However, we also found that 128 participants had a positive feeling about the study (feeling *good*, *interested*, or *confident*), mentioning their curiosity for online voting or their satisfaction regarding the security of the app. Those participants also rated the usability and UX of the application better than the others, supporting our previous idea of the benefits of a good interface.

We also note that in previous studies using the Selene protocol, for example [9, 22], the usability and user experience of Selene obtained higher scores. In these studies, Selene was implemented as a mobile app with a linear workflow, and without the faking mechanism. As a result, participants just cast their vote and verify that it was correctly recorded. In our study, all participants had to go through the faking feature, which might be a reason for participants' confusion. Further, participants could navigate through the pages without a unique workflow. The lack of linearity and the faking mechanism could also have lowered the usability score. This low score should be seen in the context of the study: we wanted to evaluate the full implementation with all participants testing the faking mechanism. In a real election it is unlikely that all voters need this feature. verification phases would probably increase the satisfaction of voters.

Finally, we hypothesize that the vote buying scenario could have led to lower trust: the qualitative feedback has shown that several people were shocked by the possibility of showing their vote to a coercer/vote buyer, and was sometimes seen as vote selling. In fact, the mechanism is designed to prevent vote buying, since a vote buyer cannot detect if it is a fake tracker. The security feature and the exacerbation of a possible threat has possibly decreased the trust from the participants, when being misunderstood. We can also note that around 50% of participants were positive to online voting and more than 90% did not have negative opinion about it before the study, adding credence to this assumption.

6.4 Recommendations

Here we provide a list of recommendations: four concern the development of future voting systems (VS), and two are about the design of user studies (US).

[VS] Focus on understandability We found that participants who understood our security features rated trust higher than other participants. However, we saw that our application was confusing and tasks too complex. When providing a new security feature one must ensure it is correctly understood to obtain an increase of trust. It is crucial to provide a transparent interface, with understandable features, to increase trust and acceptance.

[VS] Provide an easy-to-use interface While we must provide understandable and transparent information to participants, it also remains important to keep the interface as simple as possible. People who got stressed and overwhelmed by the application were less satisfied. Indeed, we found that the participants who rated the application better had a positive feeling during the study. Hence, we recommend remaining simple and straightforward, keep the workflow as linear and guiding as possible.

[VS] Raise awareness and improve education Many participants highlighted the illegality of vote buying. To them, the fact that the law is already designed to counter some threats is sufficient to trust the system. However, if a voting system is not trustworthy opens a door to attackers. We recommend communicating good practices in security and risks that could arise from a misuse of the procedure. Good education, as highlighted in previous work on mental models [34, 35] and in [13], is key to trusted applications.

[VS] Adapt the interface to the voters' profile Many participants did not see the need for a receipt-free feature (in the context of the participants' country). For future implementations, we suggest adapting the interface that will be more realistic to a targeted audience, making receipt-free aspects optional.

[US] Reduce the complexity and simplify (online) user studies We have discussed that many participants were confused during the test. We know from previous studies [21, 34] that the concept of Verifiability is hard to understand. The receipt-free feature increased the complexity. We learned that Prolific's participants need guidance to follow a study correctly, as they won't take time to explore an application. We recommend simplifying such user studies.

[US] Use the right tool In relation to limitations observed with Prolific, we further recommend in-person interviews for studies about understanding. The bias of satisficing does not help participants to focus and take time to understand the features and new concepts provided. In this study, we had a small number of participants who clearly understood the features, and we saw a correlation between their understanding and trust in the system. For an evaluation of voters' understanding and of the user experience, in-person studies with focus groups and/or interviews will bring better insights.

7 Conclusion and Future Work

In this paper, we defined trust in a voting system, and proposed a new questionnaire to assess it. We also designed and conducted a user study to evaluated the Selene voting system, including its receipt-free mechanism. Our application was tested by 300 participants; we evaluated their experience by measuring their understanding through a unique game design, and assessed their trust in the system using the new questionnaire. While the usability and trust factors were rated low in the experiments, the results supported a positive relation between trust and understanding. This let to recommendations to increase trust and usability in voting applications and to improve future user studies. Our recommendations are: 1) Focus on the understandability, 2) Provide an easy-to-use-interface, 3) Raise awareness and improve education, 4) Adapt the scenario to the audience, 5) Reduce the complexity and 6) Use the right tool. The first four apply to any (verifiable) voting system, the two last concern the execution of such trials.

For future research, it would be interesting to compare the feedback from another country, where our scenario is more common. We could set up a two-players game where one participant plays the role of a coercer or vote buyer and another plays the role of the voter, to see if the mechanism is better understood by the participants. We plan to apply and validate our trust questionnaire for other e-voting protocols and compare to [1].

Acknowldgements. The research was supported by Luxembourg National Research Fund FNR C21/IS/16221219/ImPAKT and C22/IS/ 17232062/SpaceVote, and NCBR Poland under the project SpaceVote POLLUX-XI/14/SpaceVote/2023.

References

1. Acemyan, C. Z., Kortum, P., and Oswald, F. L. The trust in voting systems (tvs) measure. *International Journal of Technology and Human Interaction (IJTHI) 18*, 1 (2022), 1–23.
2. Acemyan, C. Z., Kortum, P. T., Byrne, M. D., and Wallach, D. S. Users' mental models for three end-to-end voting systems: Helios, prêt à voter, and scantegrity II. In *Human Aspects of Information Security, Privacy, and Trust - Third International Conference, HAS 2015* (2015).
3. Agbesi, S., Dalela, A., Budurushi, J., and Kulyk, O. What will make me trust or not trust will depend upon how secure the technology is": Factors influencing trust perceptions of the use of election technologies. In *Proceedings of Seventh International Joint Conference on Electronic Voting* (2022), University of Tartu.
4. Alsadi, M., and Schneider, S. Verify my vote: Voter experience. In *Electronic Voting - E-Vote-ID 2020 (TalTech Proceedings)* (2020).
5. Bangor, A., Kortum, P. T., and Miller, J. T. An empirical evaluation of the system usability scale. *International Journal of Human–Computer Interaction 24*, 6 (2008), 574–594.
6. Bella, G., Curzon, P., Giustolisi, R., and Lenzini, G. A socio-technical methodology for the security and privacy analysis of services. In *COMPSAC Workshops* (2014), IEEE Computer Society, pp. 401–406.
7. Chiang, L. Trust and security in the e-voting system. *Electronic Government, an International Journal 6*, 4 (2009), 343–360.
8. Delaune, S., Kremer, S., and Ryan, M. Coercion-resistance and receipt-freeness in electronic voting. In *19th IEEE Computer Security Foundations Workshop, (CSFW-19)* (2006), IEEE Computer Society, pp. 28–42.
9. Distler, V., Zollinger, M.-L., Lallemand, C., Rønne, P. B., Ryan, P. Y., and Koenig, V. Security–visible, yet unseen? how displaying security mechanisms impacts user experience and perceived security. In *CHI Conference on Human Factors in Computing Systems (CHI '19)* (2019).
10. Glass, A., McGuinness, D. L., and Wolverton, M. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces* (2008), pp. 227–236.
11. Hao, F., and Ryan, P. *Real-World Electronic Voting: Design, Analysis and Deployment.* Auerbach Publications, 2016.
12. Juels, A., Catalano, D., and Jakobsson, M. Coercion-resistant electronic elections. In *ACM Workshop on Privacy in the Electronic Society* (2005).
13. Kirlappos, I., and Sasse, M. A. What usable security really means: Trusting and engaging users. In *Human Aspects of Information Security, Privacy, and Trust* (2014), vol. 8533 of *Lecture Notes in Computer Science*, Springer, pp. 69–78.
14. Kulyk, O., and Neumann, S. Human factors in coercion resistant internet voting – a review of existing solutions and open challenges. In *E-Vote-ID* (2020).
15. Kulyk, O., Neumann, S., Budurushi, J., and Volkamer, M. Nothing comes for free: How much usability can you sacrifice for security? *IEEE S&P 15*, 3 (2017).

16. Lallemand, C., and Koenig, V. Lab testing beyond usability: Challenges and recommendations for assessing user experiences. *Journal of Usability Studies* (2017).
17. Levitt, S. D., and List, J. A. What do laboratory experiments tell us about the real world. *Journal of Economic Perspectives* (2007), 153–174.
18. Llewellyn, M., Schneider, S., Xia, Z., Culnane, C., Heather, J., Ryan, P. Y. A., and Srinivasan, S. Testing voters' understanding of a security mechanism used in verifiable voting. In *2013 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE 13)* (2013).
19. Luhmann, N. *Trust and Power*, 3 ed. Polity Press, 2017.
20. Malheiros, M., Jennett, C., Seager, W., and Sasse, A. Trusting to learn: Trust and privacy issues in serious games. In *Trust and Trustworthy Computing* (2011), Springer.
21. Marky, K., Kulyk, O., Renaud, K., and Volkamer, M. What did I really vote for? On the usability of verifiable e-voting schemes. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)* (2018), ACM.
22. Marky, K., Zollinger, M.-L., Roenne, P. B., Ryan, P. Y. A., Grube, T., and Kunze, K. Investigating usability and user experience of individually verifiable internet voting schemes. *ACM Transactions on CHI 28*, 5 (2021).
23. Neto, A. S., Leite, M., Araújo, R., Mota, M. P., Neto, N. C. S., and Traoré, J. Usability considerations for coercion-resistant election systems. In *17th Brazilian Symposium on Human Factors in Computing Systems* (2018), pp. 1–10.
24. Pieters, W. Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology 13*, 1 (2010), 53–64.
25. Prolific. Prolific. `https://www.prolific.co/`.
26. Ryan, P. Y., Bismark, D., Heather, J., Schneider, S., and Xia, Z. Prêt à voter: a voter-verifiable voting system. *IEEE transactions on information forensics and security 4*, 4 (2009), 662–673.
27. Ryan, P. Y. A., Rønne, P. B., and Iovino, V. Selene: Voting with transparent verifiability and coercion-mitigation. In *Financial Crypto* (2016).
28. Ryan, P. Y. A., Schneider, S. A., and Teague, V. End-to-end verifiability in voting systems, from theory to practice. *IEEE S&P 13*, 3 (2015), 59–62.
29. Sallal, M., Schneider, S., Casey, M., Dragan, C., Dupressoir, F., Riley, L., Treharne, H., Wadsworth, J., and Wright, P. VMV: Augmenting an internet voting system with Selene verifiability. `arXiv:1912.00288`, 2019.
30. Schneider, S., Llewellyn, M., Culnane, C., Heather, J., Srinivasan, S., and Xia, Z. Focus group views on Prêt à Voter 1.0. In *2011 International Workshop on Requirements Engineering for Electronic Voting Systems* (2011).
31. Schrepp, M. User experience questionnaire handbook. `https://www.ueq-online.org/`, 2018.
32. Schrepp, M. The extended user experience questionnaire. `http://ueqplus.ueq-research.org/`, 2019.
33. Schürmann, C. Electronic elections: Trust through engineering. In *First International Workshop on Requirements Engineering for e-Voting Systems* (2009).
34. Zollinger, M., Distler, V., Rønne, P. B., Ryan, P. Y., Lallemand, C., and Koenig, V. User experience design for e-voting: How mental models align with security mechanisms. In *E-Vote-ID 2019, TalTech Proceedings* (2019).
35. Zollinger, M.-L., Estaji, E., Ryan, P. Y., and Marky, K. "Just for the sake for transparency": Exploring voter mental models of verifiability. In *Electronic Voting - Sixth International Joint Conference, E-Vote-ID 2021* (2021).