Towards improving the efficacy of code-based verification in internet voting

Oksana Kulyk
Melanie Volkamer
Monika Müller
Karen Renaud

This is the Author Accepted Manuscript of a conference paper published in Financial Cryptography and Data Security: FC 2020 International Workshops, AsiaUSEC, CoDeFi, VOTING, and WTSC, Kota Kinabalu, Malaysia, February 14, 2020, Revised Selected Papers

The final publication is available at Springer via DOI: http://dx.doi.org/10.1007/978-3-030-54455-3_21

Towards Improving the Efficacy of Code-Based Verification in Internet Voting

Oksana Kulyk[1,2], Melanie Volkamer[2], Monika Müller[2], and Karen Renaud[3,4]

[1] IT University of Copenhagen, Denmark, okku@itu.dk
[2] Karlsruhe Institute of Technology, Germany, {name.surname}@kit.edu
[3] Abertay University, Scotland, k.renaud@abertay.ac.uk
[4] Rhodes University, South Africa

Abstract. End-to-end verifiable Internet voting enables a high level of election integrity. Cast-as-intended verification, in particular, allows voters to verify that their vote has been correctly cast, even in the presence of malicious voting devices. One cast-as-intended verification approach is code-based verification, used since 2015 in legally-binding Swiss elections. We evaluated the Swiss paper-based polling sheet and voting interface, focusing on how well it supported voters in verifying their votes. We uncovered several potential issues related to manipulation detection. We improved the paper-based polling sheet and voting interface accordingly. Then, we carried out a between-subjects lab study with 128 participants to compare the original and improved sheet and interface wrt. usability and its effectiveness in supporting manipulation detection. Our improvements significantly enhanced detection. Our study delivered insights into participants' somewhat ineffectual reactions to detected anomalies, i.e. starting over again and trying to cast the same vote again, or calling the telephone number provided by the interface. This problem is likely to manifest in any verifiable voting system and thus needs to be addressed as future work.

Keywords: e-voting, verifiability, usability

1 Introduction

As the world's population increases, traditional elections become more expensive and challenging [13]. The diffusion of the Internet has changed the way we vote [24]. While some of these changes have positively impacted our lives, there are negative side effects too due to the activities of malicious actors, such as dissidents seeking to disrupt Internet-enabled elections [10, 12]. Attackers could replace or discard votes by manipulating vote casting devices (laptop or smartphone) or the vote-casting software. This causes problems *first*, for the election authorities, who need to ensure the integrity of the election and *second*, for the voters themselves, who need to have confidence in the election outcome.

One way of reassuring both stakeholders, and of maximising the probability that malicious activities will be revealed, is to build verifiability into the voting system allowing voters to verify that their vote has indeed been cast as they intended. Voters can all help to reveal anomalies resulting from malicious activities.

However, for verifiability to deliver this assurance, it relies on two key assumptions: (1) that voters will indeed perform all the necessary verification steps, and (2) that they will do so correctly. These assumptions have been challenged by empirical studies into verifiable Internet voting systems [1, 14, 20, 27]. The complexity and unfamiliarity of verification process can prevent voters from performing the necessary steps correctly, or at all.

If voters do not verify correctly, this renders the integration of verifiability into Internet voting systems fruitless, particularly as ever more sophisticated attacks emerge. If malicious actors can control what voters see, by manipulating the vote-casting device and/or vote-casting software, they can subvert the process. They could, for example, assure the voter that their vote has been cast successfully although, in reality, additional steps are mandated to finalise the process. Malicious attacks have been discussed in the literature [11, 17], but we are not aware of any empirical evaluation of the efficacy of verification provision being carried out. Nor have researchers systematically addressed these issues. Both gaps are narrowed by our research.

The widely-used Swiss verifiable Internet voting system utilises code-based verification (Section 2) so we commenced by investigating its usability and the ease of cast-as-intended verification. We then developed and improved the verification mechanism (Section 3) and evaluated both the original and improved mechanisms in a between-subjects lab study with 128 participants (Section 4). We found that the improvements preserved usability and participants detected significantly more manipulations (Section 5). We discuss our findings in Section 6, including lessons learned related to those aspects influencing 'correct' verification. We conclude in Section 7.

2 Background

We describe the idea and the process of the code-based verification, including its implementation in the Swiss system, and provide further background by describing related work on the usability of cast-as-intended verification.

2.1 Code-Based Verification

Code-based verification systems issue voters with a unique *polling sheet*, delivered via snail mail. This provides the voting system's website address, instructions, and one or more codes to facilitate verification. A number of variants of code-based verification approaches exist, including the one used in Swiss elections [25]. Their code sheets provide three types of codes: (1) verification[5], (2) confirmation *and* (3) finalisation. The system responds to a cast vote by displaying a 'verification code'. The voter compares this to the code that appears on their personalized code sheet. If they match, the voter enters the confirmation code (also provided on the sheet). Otherwise, they ought to report the anomaly to the election authorities. A finalisation code is subsequently displayed to reassure the voter that the voting system has indeed cast their vote as intended, and that the voter has confirmed that the code matching their choice was correctly displayed.

[5] This type of code is also commonly referred to as a check or return code in the literature.

Note that the code-based verification is not sufficient to ensure the election integrity on its own; namely, one still has to ensure that the election authorities are trustworthy and perform the tallying procedure correctly. Further mechanisms, such as tallied-as-stored verifiability methods, should be employed for this purpose.

2.2 Manipulations

Achieving verifiability relies on voters assiduously and attentively going through all the verification steps. If an adversary is able prevent the voter from verifying correctly, providing verifiability fails to achieve its aim. An adversary might manipulate the voting interface and subvert the verification protocol (see [17]).

An adversary wanting to replace a vote for candidate 'A' with a vote for candidate 'B' might deploy various strategies: (1) Replace the verification code with another code, or (2) Remove the verification code entirely, or Both strategies can be applied by keeping the rest of the interface as is or by (3) adding messages such as: "*thank you for your vote - you are done*" or "*the verification code for your candidate is correct*" to allay suspicions. Aside from these two manipulations, the adversary can take other steps in ensuring that the verification procedure is not followed correctly, such as withholding the confirmation code. This attack type, however, leaves a trace in the voting system, when the vote is recorded as "attempted" *but not confirmed*. A large number of these might raise alarms during auditing. Furthermore, even if the attack succeeds, the adversary only blocks the vote, but is unable to replace it. Note, however, that in order to account for all the attacks exploiting this vulnerability, a systematic investigation involving attack trees is needed. Even then, achieving comprehensiveness remains an open challenge when it comes to exploiting the human factor.

2.3 Related Work

Research into *verification-related mental models* [21, 23] revealed a number of factors that could prevent voters from verifying, such as a lack of knowledge, required effort and misconceptions.

Other studies focused on *usability* evaluating user experiences and voter satisfaction [9, 16, 18, 19, 22, 28]. Some studies reported high satisfaction, while others uncovered usability issues. A study into the usability of the Norwegian Internet voting system, which relies on code-based verification, mentions a lack of understandability related to the range of different codes. None of these studies measured verification effectiveness i.e. whether their participants were able to verify their votes. This was indeed evaluated for verifiable voting systems, such as for *Prêt à Voter* and Scantegrity II in [1, 2], for BingoVote in [5], for StarVote in [3], for EasyVote in [7, 8], for the ballot-marking devices used in the US elections in some of the states [6] and for Helios Internet voting system in [1, 14, 20, 27]. Some reported high rates of verification effectiveness [3, 7], others reported issues [1, 1, 2, 5, 20] including verification misconceptions, which resulted in participants being unable to verify their votes successfully.

Some studies evaluated the *effectiveness of code-based verification* by deliberately introducing manipulations during the process. Kulyk *et al.* [15] evaluated the effectiveness of a code-based approach and manipulated the verification code, replacing it

with an incorrect code. While all participants detected the manipulation, the removal of the verification code was not tested. The study focused exclusively on verification effectiveness, without evaluating other usability aspects, such as satisfaction or efficiency. The study by Gjøsteen and Lund [11] evaluated the Norwegian Internet voting system, which provided verifiability by sending the verification codes to the voters via SMS after they had submitted their vote. In the attacks simulated in the study, no such code was sent, but only 6 of 30 participants detected this. Similar to [15], the voting interface was unaffected by the manipulations. No specific evaluation has been carried out to detect whether voters detect user interface manipulations while verifying their cast votes.

3 Improving the Swiss Voting System

In this section, we describe our initial analysis of the Swiss voting system as well as our modifications calculated to improve the usability of their cast-as-intended verification, to make it more likely that voters will detect manipulations to their device's voting interface.

3.1 Issues with the Original System's Cast-as-Intended Verification

We organized a brainstorming session between the authors and also arranged feedback sessions with other participants including a lay person, an expert in human-computer interaction, an expert in general security, and an expert in electronic voting security. Participants received background information and were instructed to think aloud while casting a vote for a specific candidate. Afterwards they were given the election information sheet, including the polling sheet providing the codes. They used our laptop to interact with the voting system. Notes were taken and the session ended by eliciting responses about what would prevent them from verifying their votes.

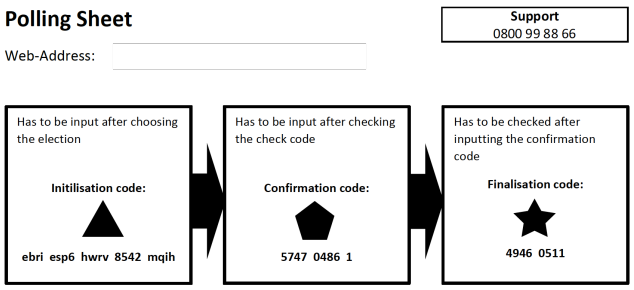

Fig. 1: The version of the polling sheet

The sessions revealed a number of issues. Here, we focus on those relevant to cast-as-intended verification steps (note: the remaining general usability issues were addressed in producing our improved version).

Lack of a Step-by-Step Process The process of casting and verifying a vote consists of a number of steps that are new to the voters, given their familiarity with traditional paper-based voting. The polling sheet makes it look as if there are three steps (three boxes linked by arrows): entering the initialisation code, entering the confirmation code, checking for the finalisation code. This is a serious issue because the vital step of comparing the displayed code with the matching vote choice code could easily be omitted. The point at which the vote is finally stored, and the process concluded, is unclear. Voters might assume that once the confirmation code has been entered, their vote has been cast. They might not notice that the finalisation code is incorrect or missing. The voter's interaction with the system requires them to conduct more steps than those communicated on the original polling sheet and the steps do not match. This makes it even less likely that the voter will notice manipulations. It is also not clear what the voter ought to do if codes do not match. Instructions provided by the interface cannot be trusted because these can be removed. Details about who to contact for voter support is not prominent on the polling sheet (and it is not clear that support should be contacted if the voter detects a code mismatch)

Unclear Explanations The presentation of so many codes, without explanation, is confusing. Voters could easily be left wondering why they are needed and why they ought to be verified. Moreover, no instructions are provided to tell voters what to do if the codes fail to appear on the interface. The same term is used for different concepts. For example, the initialisation code is actually an authentication code, while the others are codes used for verification. Finally, voters are not told which codes ought to be entered into the system, and which ought to be compared to codes on the sheet (but not provided to the system).

Finalisation Page Header While the voters need to compare the displayed finalisation code with their sheet's code in order to make sure that their vote has indeed been cast, the final finalisation code display page includes a "thank you for your vote" message, which might mean that voters assume they are done, and unwittingly omit the final verification step.

These issues point to the lack of clarity. A voter who is unfamiliar with the process might miss crucial verification steps, e.g. forgetting to compare the verification codes. This requirement is not explicitly mentioned on the sheet rendering the voting system vulnerable to manipulations described in Section 2.2.

3.2 Proposed Improvements

Based on the aforementioned feedback, we proposed improvements to the sheet and the voting interface. We focused primarily on improving the voting materials, since, as discussed in Section 2, an adversary who controls the voting environment is likely also to be capable of modifying the website interface. The improvements were refined over several feedback sessions:

Polling Sheet The layout of the text on the provided sheet was changed in order to provide a more structured overview of the steps the voter has to perform to cast and verify her vote. In the first place, we included a sequence diagram with the individual steps clearly marked, including alternative actions to be taken if verification reveals potentially malicious activities. We also rewrote the explanation texts to improve understandability, and referred to the initialisation code as a "password" in order the better to distinguish it from the codes used during the verification process. The resulting polling card is provided in Figure 2.

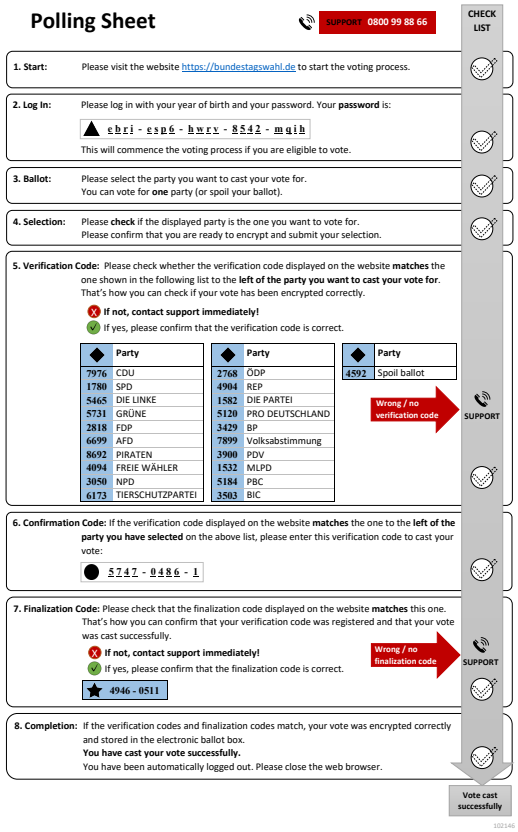

Fig. 2: The improved version of the polling sheet (translated from German). Both the website and the phone number are fictitious.

Voting Interface We rewrote the explanatory texts to improve clarity and to provide the voters with more explicit guidelines regarding the vote casting and the verification processes. An example of the resulting interface is provided in Figure 3.

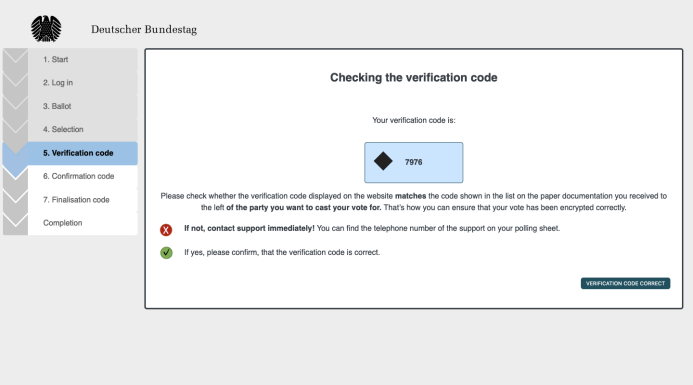

Fig. 3: The improved version of the verification web page.

4 Evaluation Methodology

We describe the methodology we followed in evaluating our proposed improvements and their comparison with the original Swiss system[6].

4.1 Manipulations

We evaluate the voter's ability to detect two possible manipulations: *replacing* the verification code, or *removing* it from the output screen. We made the following modifications to the voting website interface:

Replace We removed all the instructions in the interface that explained what to do if the verification code did not match the one on the code sheet. We also modified the instructions to verify the verification code by mentioning that the code could be found anywhere on the code sheet. We added a "Support" button, which gave the voter a number to call if they had any issues. The number was different from the one on the polling sheet, assuming that an attacker would replace it to prevent voters from reporting it to the authorities.

Remove We replaced the output of the verification code and the verification instruction with a message congratulating the voter on casting their vote.

4.2 Evaluated Metrics and Hypotheses

The main focus of our study is to evaluate how well the participants were able to detect both *replace* and *remove* type manipulations in both the original and improved systems. Because the aim of our improvements was to increase the manipulation detection rate, we tested the following hypotheses:

[6] For screenshots of the voting website and polling sheets for both of the evaluated systems, see https://secuso.org/code-based-supplemental-material

$H_{replace}$ Participant voters are more likely to detect manipulations that replace the verification code when using the improved system than when using the original Swiss system (the *replace*-Manipulation)..

H_{remove} Participant voters are more likely to detect manipulations that remove the verification code when using the improved system than when using the original Swiss system (the *remove*-Manipulation).

We assessed the usability of both systems under normal conditions, that is, in absence of manipulations. We therefore evaluate the following metrics:

Efficiency How long did participants take to cast their votes and to complete verification in absence of manipulations.

Effectiveness How many participants are able to cast their votes in absence of manipulations.

Satisfaction Participants' satisfaction, using the SUS measurement tool (calculated on a scale from 0 to 100).

Because our improvements focused on enhancing manipulation detection, we did not expect any differences in general usability. As such, while we aimed to make the voting process clear by adding instructions, the presence of these instructions, while improving understandability, could decrease efficiency if they take longer to read. We nonetheless measure these criteria and report on descriptive statistics, in order to identify opportunities for further refinements. We furthermore conducted a qualitative evaluation of the usability of both systems, by analysing the feedback from the participants who used the unmanipulated systems.

4.3 Study Type

We conducted a between-subjects lab study, with participants randomly assigned to one of the following six groups:

Original-No-Manipulation The group interacting with the original system interfaces, with no manipulation occurring,

Original-Replace The group interacting with the original system interfaces and subjected to a manipulation that *replaces* the verification code,

Original-Remove The group interacting with the original system interfaces and subjected to a manipulation that *removes* the verification code,

Improved-No-Manipulation The group interacting with the improved system interfaces, with no manipulation occurring,

Improved-Replace The group interacting with the improved system interfaces and subjected to a manipulation that *replaces* the verification code,

Improved-Remove The group interacting with the improved system interfaces and subjected to a manipulation that *removes* the verification code,

The purpose of using six groups is to support evaluation of the usability of both the original system and the improved variant under two conditions: normal voting, where no manipulation occurs, and an attempted attack with the adversary either replacing

or removing the verification code. The two groups without manipulation *Original-No-Manipulation* and *Improved-No-Manipulation* are used to study the system under normal conditions. We evaluate user satisfaction and effectiveness in terms of being able to cast and verify votes in the absence of manipulations. The four groups with manipulations (*Original-Remove*, *Original-Replace*, *Improved-Remove* and *Improved-Replace*) allow us to evaluate the effectiveness of the system under the attempted attack scenario, namely, the ability of participants to detect different kinds of manipulations with either the original system (hypothesis $H_{replace}$) or the improved system (hypothesis H_{remove}).

4.4 Study Procedure

The participants in each group are told that the evaluation goal was to study the usability of an Internet voting system, to be used in forthcoming German Parliament elections. Before the study, the participants were asked to read and sign a consent form, detailing the study procedure (not mentioning the manipulations), explaining that data collected during the study was anonymised and would be analysed by the research group which the paper authors are a part of. They were told that they could abort the study at any time, in which case all the data collected so far would be deleted. The participants were told about the reimbursement of 10 Euros that they could claim for participating.

After signing the consent form, the participants were presented with a role card telling them to assume a role of Mr. or Ms. Müller, born in 1970, opting to cast their election vote over the Internet. The role card informed them that they would receive the election materials to be used to cast a vote for the SPD[7] party[8]. The participants were provided with a mock-up welcome letter from the election authorities with general information about the election and instructions to look for the information necessary for using the Internet voting system (on a polling card, also handed out to them at the beginning of the study.

As soon as the participants indicated that they were finished reading the instructions, they were asked to use the voting system installed on the lab computer. The system consists of the mock-ups of the interfaces for the corresponding system (that is, either the original or the improved interfaces)[9]. The mock-ups simulate the German Parliament election (to be conducted in 2021) with the list of candidates from the 2017 election. The participants were instructed to cast a vote for a party outlined on their role card. For the participants in *Original-No-Manipulation*, *Improved-No-Manipulation* groups, no manipulation took place, so that voters were able successfully to complete the process of casting, verifying and confirming their votes. The time they took to complete casting the vote was noted by the examiner. The participants in the remaining groups were subjected to manipulation depending on their group – that is, the verification code was either removed or replaced, depending on experimental condition.

If a participant noticed the manipulation, they were asked how they would behave if this occurred in an actual election. If the participant answered that they would call the support number, they were asked whether they would use the number on the website, or

[7] Germany's Social Democratic Party

[8] Participants were asked to cast a vote for a specific party to preserve ballot secrecy.

[9] Cast votes were neither stored nor processed

the number given on the polling sheet. Afterwards, they were debriefed and told about the real purpose of the study. All participants were asked to fill in questionnaires assessing their satisfaction with the system, as well as being asked questions about whether they had issues with casting their votes and what they found positive or negative regarding the voting website and the polling sheet, participants who did not notice manipulations, as well as the participants who were not subjected to manipulations, were debriefed about the real purpose of the study.

4.5 Recruitment and Ethics

The participants were recruited using snowball sampling. They were told that the study would take around 25 minutes and were offered a reimbursement of 10 Euros, which is above the minimum hourly German wage (around 9 Euro). The authors' institutional ethical and data protection guidelines were followed.

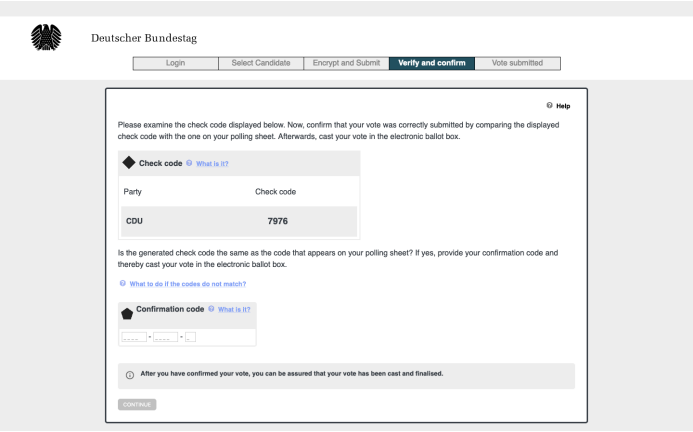

Fig. 4: The version of the verification web page, modeled after the original Swiss system and adjusted to resemble the German election scenario (translated from German).

5 Evaluation Results

144 participants took part. 16 were removed prior to the analysis due to deviations in the study procedure, such as the examiner handing out the materials for the wrong group by mistake. Of the remaining 128 participants, aged 20-81, with the mean age of 34.34 and standard deviation of 15.54. 66 were female and 62 male.

5.1 Manipulation Detection

In order to evaluate the hypotheses provided in Section 4.2, we consider the number of participants who detected a manipulation presented to them during the study (that is, either a replaced or an omitted verification code). We performed a comparison between

the groups *Original-Replace* and *Improved-Replace*, as well as between the groups *Original-Remove* and *Improved-Remove*. An overview of the numbers of participants in each group who detected the corresponding manipulation is provided in Table 1. The results for both manipulations are analysed using a one-sided Fisher's exact test.

Replacing the Verification Code While the majority of the participants in the *Original-Replace* group detected the manipulation, one fourth failed to do so. On the other hand, all participants in the *Improved-Replace* group detected the verification code replacement. $H_{replace}$ is therefore confirmed ($p = 0.0187$, odds ratio 95% CI: [0, 0.662]).

Reaction to Detected Misaligments We asked all of the participants who detected the manipulation what they would do if they had such experience in real-world election. Some of the participants have noted that they are likely to login again and try one more time to cast the vote. An attacker can take advantage of such behaviour, for example, by trying to manipulate the vote during the voter's first attempt and leaving the vote intact if the voter tries again, thus ensuring that the manipulation remains unreported to the election authorities. Some participants further mentioned that they would call the support. However, when asked, which number they would use, several said they would call the number they saw on the website. As mentioned in Section 4, this number had also been altered under the assumption that an adversary would probably display a fake number on the voting interface to reassure concerned voters. Voters who call this number would likely reach the adversary him or herself, and not the election authorities.

	detected	undetected		detected	un detected
Original-Replace	16 (76.2%)	5 (23.8%)	Original-Remove	2 (9.1%)	18 (90.9%)
Improved-Replace	23 (100%)	0	Improved-Remove	9 (43.48%)	12 (56.52 %)

Table 1: Number of participants detecting and not detecting the manipulation of replacing or removing the verification code

Removing the Verification Code The manipulation that involves removing the verification code was particularly hard for participants to detect. As such, only two of 20 participants managed to do so in the group using the original system. The results were better in the group using the improved system. Even so, more than half of these participants (12 out of 21) did not detect manipulation either. H_{remove} is therefore confirmed ($p = 0.02$, odds ratio 95% CI: [0, 0.752]). As there was no "Support" Button on this particular screen, as opposed to the *Replace*-Manipulation, we did not ask participants who they would call.

5.2 General Usability

As described in Section 4.2, we also consider general usability of the both systems in the absence of manipulations.

Effectiveness We looked whether the participants that were not subjected the manipulation were able to cast their votes successfully. Of the 47 participants in groups *Original-No-Manipulation* and *Improved-No-Manipulation*, only one was not able to complete the vote casting process. The participant was interacting with the original system and thought that the displayed verification code was incorrect, resulting in a false positive result during verification.

Satisfaction We compared the SUS scores of the *Original-No-Manipulation* and *Improved-No-Manipulation* groups (i.e those who experienced the system without being disrupted by a manipulation). We used only the scores from those who completed the vote casting process correctly. Both systems were awarded a high score (an average of 79.9 from 22 participants for the original system and an average of 80.9 from 20 participants for the improved), which is classified as "good" according to the scale proposed by Bangor *et al.* [4].

Efficiency We measured the time it took participants from commencing the voting process to finalising their cast vote in absence of manipulation (i.e. groups *Original-No-Manipulation* and *Improved-No-Manipulation*). On average, the participants required 175.86 seconds using the original system and 180.35 seconds using the improved system.

5.3 Qualitative Feedback

In order to identify further issues and potential improvements of the system, we considered the answers from the participants not subjected to manipulation to the following questions:

- Did you experience any issues with casting your vote? Which ones?
- What did you find positive about the polling sheet?
- What did you find positive about the voting website?
- What did you find negative about the polling sheet?
- What did you find negative about the voting website?

We summarised the responses to these questions for each one of the two systems below, providing the number of participants who mentioned each answer while omitting these numbers if an answer was only mentioned by either one or two participants.

Original System The majority of the participants using the original system (18 out of 24) did not name any issues they had with casting their votes. The issues named by the rest of participants were related to the amount of codes, their complexity and difficulties in entering them without making errors, inability to distinguish between similar-looking characters in the codes (namely, i and l) and the font size being too small.

When asked about what they considered positive about the polling sheet, the participants mentioned the clarity of the instructions (11) and the comprehensibility of the polling sheet (7). Others commented on feeling secure due to assurance via different

codes, the symbols for the codes being helpful, the sheet being compact, the presence of an emergency number and the choice of the headers. Five participants did not comment.

With respect to positive feedback on the voting website, the participants mostly commented on the system being fast (12) and easy to use (11). Further comments were related to the convenience of being able to cast their vote over the Internet (5), feeling secure in casting the vote, clarity of the instructions and the possibility of decreasing paper waste by using Internet voting. One participant did not give any positive feedback.

When asked about what the participants did not like about the polling sheet, the most common issues were the lack of information about the codes on the sheet, making the codes confusing without seeing the website (named by 7 of 24 participants) and the font being too small (named by 4 participants). Other issues were the complexity of the codes, too much information packed on the sheet, wanting more information about the technical aspects of the system, wanting to see a second polling sheet that outlines an example of how to cast a vote, not liking the use of the word “Support", finding the identification step illogical, general criticisms of the instructions and finding the term “verification code" confusing. Seven of 24 participants did not provide any negative feedback.

When asked for negative feedback about the voting website, five of 24 participants had doubts about the security of the system, such as a lack of control of whether the polling sheet is actually used by an authorised voter, the influence of third parties on the election outcome, or a general feeling of insecurity. Other issues were the design of the website looking untrustworthy, difficulties in navigating the help page, insufficient feedback when the vote was cast successfully, small font, lack of instructions regarding how many parties one is allowed to choose, lack of information about the parties on the ballot, finding the identification step illogical and feeling that one would miss the traditional aspects of voting, such as walking to the polling booth. Nine participants did not provide any negative feedback.

Improved System Most of the participants using the improved system (20 out of 22) did not experience any issues during vote casting. The rest mentioned the overall inconvenience of the process, and the irritation with the system deleting entries after the Enter button was pressed.

When asked about positive feedback on the voting sheet, 11 of 22 participants commented on the enhanced comprehensibility and 5 mentioned the clear structure of the sheet. Three liked the check list on the right of the polling sheet, and the rest liked the use of color and the length of the polling sheet. Three did not give any positive feedback.

Similar to the original system, the most commonly named positive aspects of the voting website were the ability to cast a vote online (11 of 22 participants), the system being fast (5) and easy to use (10). Other positive mentions were the instructions for the system, overall clarity of the interface and the possibility of reducing paper waste.

Most of the participants did not mention any negative issues with the polling sheet (14 of 22). The remaining participants mentioned issues such as the complexity of writing down the codes, the large number of steps and the overall complexity of the procedure, design choices such as colors used in the table, the length of the instructions and the need to read them to avoid errors, and feeling that the instructions were redundant.

Among the negative issues with the voting website, the participants mentioned the complexity of the system, the inconvenience of having to check the codes, the number of steps required to traverse and the number of codes to enter, wanting a better user interface design (e.g. finding the party list too long), wanting a better understanding of the security that the system provides, general concerns with Internet voting and missing the social aspects of polling-station voting.

6 Discussion

Our study has shown that participants struggle to detect manipulations if an adversary manages to manipulate the voting interface. This was particularly noticeable when the verification code was removed. Even after improvements, fewer than half of the participants detected this particular manipulation; only two detecting it using the original system. The detection rates are even worse than those reported by Gjøsteen and Lund [11], where a fifth of the participants (6 of 30) detected a missing verification code that was supposed to be received via SMS. This is possibly because the manipulation we tested involved modifying interfaces, which fooled participants into believing that all was well. This demonstrates that even verifiable voting systems remain at high risk of undetected vote manipulations. Yet an attacker could prevent the voters from completing any or all verification steps. It follows that usability (both of the systems in our study received high SUS scores, and many participants commented on their ease of use) is not the only factor that has to be considered in designing this kind of systems, confirming previous findings [17].

Future research into the design of these systems is needed in order to improve the manipulation detection rate and to address the issues related to participants' reaction to detected misaligments. Such research could focus on finding new ways to present the information about the proper voting procedure to the participants in different formats. This may include designing an information flyer with examples of correct and incorrect voting procedures, or using an interactive app to guide the voters.

Some users were unable to detect the manipulation to the verification code in the original system. This observation is different from the results of the study in [15], where all the participants were able to detect the manipulation using a code-based verification similar to the Swiss system. One possible explanation might be that the participants in the previous study were explicitly instructed to verify their vote, whereas in our study the focus was on casting a vote. A further explanation might be the intervening changes to the design of the Swiss polling sheet, so that the two studies tested different systems (actually in particular different polling sheets).

Although both the original and improved systems received high SUS scores, the complexity of the procedure and the codes was an issue, again confirming the findings from [15]. Voters might well be willing to accept complex systems if they are told that the complexity is necessary for security (see [16]). Furthermore, the voters who are informed about security protection mechanisms, such as verification, might trust the voting system more [26]. It is therefore worth investigating, whether including additional information about how the codes bolster security would be helpful to voters, and to find the best

way of providing this information without overwhelming them. This might also help to address the issues we detected with participants reaction to detected misaligments.

Our study's sample consisted mostly of younger participants. This is not representative of the voting population, but if younger people, who are more comfortable with technology, fail to detect manipulations, the issue might be even more critical for older voters. We also note that Internet voting is usually implemented as an optional voting channel, available *in addition to* traditional paper-based voting, so that voters who are not confident in their ability to use technology can still cast their votes at polling stations. Nonetheless, investigating the human factors of Internet voting with an older sample remains an important direction for future work.

The study has the common limitations of studies that measure the usability of verification, in terms of differences between the lab and the real-life behaviour. As such, real-world verification might deliver different performances, perhaps because participants are more likely to read and follow the instructions when they know that they are being observed. On the other hand, they might be more incentivised to verify their vote in a real election, and therefore to pay more attention to the verification procedure and output codes, as the integrity of their votes is more important than in the lab setting. Still, conducting a study that involves introducing vote manipulations in a real-world election would pose critical ethical and legal issues, potentially undermining the participants' trust in the election authorities. A possible middle ground can be found, for example, by conducting remote studies where the participants are not directly observed during vote casting. Another way would be to conduct mock elections without telling the participants about the real purpose of the study before they cast their votes. However, one needs to choose a topic that participants will care about (otherwise they would not be incentivised to verify), yet, manipulation of votes on such a topic will most likely trigger an emotional response that will endure even after the debriefing. Given these considerations, the obstacles to in-the-wild testing seem almost insurmountable.

Finally, we focused on two kinds of attacks that, if successful, could jeopardize election integrity by allowing an adversary to replace cast votes with votes for a candidate of their choosing or merely reducing the number of votes that go to a candidate they do not approve of. Our study clearly does not attempt to improve verification in the face of all possible manipulation tactics, especially if one takes social engineering attacks into consideration. Investigating the scope and potential success of other attacks is an important and promising direction for future work.

7 Conclusion

Internet voting systems are a relatively new innovation in the history of democracy. To reassure voters, many systems build verifiability into the systems. However, achieving verifiability requires participation from the voters themselves, who now have to carry out extra steps in addition to casting their votes. The entire concept of verifiability stands or falls based on their ability to do this, and to spot any anomalies that manifest. If voters are able to do this, nefarious activities will be uncovered. The study reported here is the first to test whether voters can detect manipulations to the voting interface when voters verify using the code-based verification approach. To maximise the chances that people

would spot manipulations, we first improved the paper-based instructions provided to walk people through the required steps. We then carried out a lab-based study with both the original and improved systems, While our refinements improved detection rates, participant voters did not universally detect the manipulations. There is clearly room for further refinements. One additional finding - which is likely to hold for any verifiable voting system - is that it is not enough to make people detect a manipulation if they then call the malicious support hotline or simply try again. This needs to be addressed as future work. What our investigation highlights is the need to consider the human in the loop when designing user interactions, especially where tasks are unfamiliar and different from the traditional way of doing things.

References

1. Claudia Z Acemyan, Philip Kortum, Michael D Byrne, and Dan S Wallach. 2014. Usability of voter verifiable, end-to-end voting systems: Baseline data for Helios, Prêt à Voter, and Scantegrity II. *The USENIX Journal of Election Technology and Systems* 2, 3 (2014), 26–56.
2. Claudia Z Acemyan, Philip Kortum, Michael D Byrne, and Dan S Wallach. 2015. From error to error: Why voters could not cast a ballot and verify their vote with Helios, Prêt à Voter, and Scantegrity II. *USENIX Journal of Election Technology and Systems (JETS)* 3, 2 (2015), 1–19.
3. Claudia Ziegler Acemyan, Philip Kortum, Michael D Byrne, and Dan S Wallach. 2018. Summative Usability Assessments of STAR-Vote: A Cryptographically Secure e2e Voting System That Has Been Empirically Proven to Be Easy to Use. *Human Factors* (2018), 1–24.
4. Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4, 3 (2009), 114–123.
5. Michael Bär, Christian Henrich, Jörn Müller-Quade, Stefan Röhrich, and Carmen Stüber. 2008. Real world experiences with bingo voting and a comparison of usability. In *IAVoSS Workshop On Trustworthy Elections (WOTE 2008)*. Leuven, Belgium.
6. Matthew Bernhard, Allison McDonald, Henry Meng, Jensen Hwa, Nakul Bajaj, Kevin Chang, and J Alex Halderman. 2020. Can Voters Detect Malicious Manipulation of Ballot Marking Devices?. In *To appear in Proc. 41st IEEE Symposium on Security and Privacy (Oakland'20)*.
7. Jurlind Budurushi, Karen Renaud, Melanie Volkamer, and Marcel Woide. 2016. An Investigation into the Usability of Electronic Voting Systems for Complex Elections. *Annals of Telecommunications* 71, 7-8 (2016), 309–322.
8. Jurlind Budurushi, Marcel Woide, and Melanie Volkamer. 2014. Introducing Precautionary Behavior by Temporal Diversion of Voter Attention from Casting to Verifying their Vote. In *Workshop on Usable Security, USEC 2014, San Diego, California, 23.02.2014*. Internet Society, Reston, VA.
9. Verena Distler, Marie-Laure Zollinger, Carine Lallemand, Peter Roenne, Peter Ryan, and Vincent Koenig. 2019. Security–Visible, Yet Unseen? How Displaying Security Mechanisms Impacts User Experience and Perceived Security. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI2019)*. Glasgow, Scotland, 605:1–605:13.
10. Martin Giles. 2019. US elections are still far too vulnerable to attack - at every level. (2019). 6 June https://www.technologyreview.com/s/613635/us-elections-are-still-far-too-vulnerable-to-attackat-every-level/, Accessed 23 June 2019.
11. Kristian Gjøsteen and Anders Smedstuen Lund. 2016. An experiment on the security of the Norwegian electronic voting protocol. *Annals of Telecommunications* 71, 7-8 (2016), 299–307.

12. J Alex Halderman. 2016. Practical attacks on real-world e-voting. In *Real-World Electronic Voting. Design, Analysis and Deployment*, Feng Hao and Peter Y. A. Ryan (Eds.). Auerbach Publications, Boca Raton, USA, 143–170.
13. Masataka Harada and Daniel M Smith. 2014. You have to pay to play: Candidate and party responses to the high cost of elections in Japan. *Electoral Studies* 36 (2014), 51–64.
14. Fatih Karayumak, Maina M. Olembo, Michaela Kauer, and Melanie Volkamer. 2011. Usability Analysis of Helios - An Open Source Verifiable Remote Electronic Voting System. In *Proceedings of the 2011 Conference on Electronic Voting Technology/Workshop on Trustworthy Elections (EVT/WOTE'11)*. USENIX Association, Berkeley, CA, USA.
15. Oksana Kulyk, Jan Henzel, Karen Renaud, and Melanie Volkamer. 2019. Comparing "Challenge-Based" and "Code-Based" Internet Voting Verification Implementations. In *IFIP Conference on Human-Computer Interaction*. Springer, Paphos, Cyprus, 519–538.
16. Oksana Kulyk, Stephan Neumann, Jurlind Budurushi, and Melanie Volkamer. 2017. Nothing Comes for Free: How Much Usability Can You Sacrifice for Security? *IEEE Security & Privacy* 15, 3 (2017), 24–29.
17. Oksana Kulyk and Melanie Volkamer. 2018. Usability is not Enough: Lessons Learned from Human Factors in Security - Research for Verifiability. *E-Vote-ID* (2018), 66–81.
18. Damien MacNamara, Paul Gibson, and Ken Oakley. 2012. A preliminary study on a DualVote and Prêt à Voter hybrid system. In *CeDEM 12 Conference for E-Democracy and Open Government, 3-4 May*. Edition-Donau-Univ. Krems, Danube-University Krems, Austria, 77.
19. Damien MacNamara, Ted Scully, and Paul Gibson. 2011. DualVote Addressing Usability and Verifiability Issues in Electronic Voting Systems. (2011). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.399.7284.
20. Karola Marky, Oksana Kulyk, Karen Renaud, and Melanie Volkamer. 2018. What Did I Really Vote For?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal, Canada, 176.
21. M Maina Olembo, Karen Renaud, Steffen Bartsch, and Melanie Volkamer. 2014. Voter, what message will motivate you to verify your vote. In *Workshop on Usable Security, USEC*. Okinawa, Japan.
22. Anne-Marie Oostveen and Peter Van den Besselaar. 2009. Users' experiences with e-voting: A comparative case study. *Journal of Electronic Governance* 2, 4 (2009).
23. Steve Schneider, Morgan Llewellyn, Chris Culnane, James Heather, Sriramkrishnan Srinivasan, and Zhe Xia. 2011. Focus group views on Prêt à Voter 1.0. In *International Workshop on Requirements Engineering for Electronic Voting Systems (REVOTE),*. IEEE, Trento, Italy, 56–65.
24. Eva Johanna Schweitzer and Steffen Albrecht. 2011. Das Internet im Wahlkampf: Eine Einführung. In *Das Internet im Wahlkampf*, Eva Johanna Schweitzer and Steffen Albrecht (Eds.). Springer, Heidelberg, Germany, 9–65.
25. Uwe Serdult, Micha Germann, Fernando Mendez, Alicia Portenier, and Christoph Wellig. 2015. Fifteen Years of Internet Voting in Switzerland [History, Governance and Use]. In *2nd International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, Quito, Ecuador, 126–132.
26. Mihkel Solvak and Robert Krimmer. 2019. The curse of knowledge. *E-Vote-ID* (2019).
27. Janna-Lynn Weber and Urs Hengartner. 2009. Usability Study of the Open Audit Voting System Helios. http://www.jannaweber.com/wpcontent/uploads/2009/09/858Helios.pdf. (2009). [Accessed: 22-December-2017].
28. Marco Winckler, Regina Bernhaupt, Philippe Palanque, David Lundin, Kieran Leach, Peter Ryan, Eugenio Alberdi, and Lorenzo Strigini. 2009. Assessing the Usability of Open Verifiable E-Voting Systems: a Trial with the System Prêt à Voter. In *Proceedings of ICE-GOV*. 281–296.