How to Assess the Usability Metrics of E-Voting Schemes

Karola Marky[1][0000−0001−7129−9642], Marie-Laure Zollinger[2], Markus Funk[1][0000−0002−8127−0592], Peter Y. A. Ryan[2], and Max Mühlhäuser[1][0000−0003−4713−5327]

[1] Telecooperation Lab - Technische Universität Darmstadt, Germany
surname@tk.tu-darmstadt.de
[2] University of Luxembourg {marie-laure.zollinger,peter.ryan}@uni.lu

Abstract. Voters play an important role in end-to-end verifiable e-voting schemes, because the schemes encourage them to carry out several security-critical tasks by themselves. If the voters cannot complete the tasks by themselves or experience a bad usability while executing them, vote manipulations by either a faulty software or deliberate attacks cannot be detected which renders verification useless. Therefore, the scheme's usability is of crucial importance and demands an early investigation of human factors when implementing e-voting systems. In this paper we give an overview of user study design challenges when investigating end-to-end verifiable e-voting schemes. We provide guidelines that address these challenges and support researchers in the design of user studies. The guidelines are based on the literature and the authors' experiences.

Keywords: E-Voting · Usability Evaluation · End-to-End Verifiability.

1 Introduction

Vote integrity means that an election's result must accurately reflect the voters' true intentions. *End-to-end verifiability* [12] is a measure for vote integrity and provides means for the voters to verify that their intentions are accurately represented in the election's result. Implementing end-to-end verifiable e-voting schemes constitutes a particular challenge due to competing security requirements that have to be assured simultaneously. A particular opponent of verifiability is *vote privacy* meaning that the voting system does not provide more evidence about the intention of specific voters than the election result does [37].

In order to maintain the vote privacy in an end-to-end verifiable e-voting scheme, the voters have to carry out several tasks by themselves. Furthermore, they have to determine the result of the verification, i.e. whether a vote is manipulated or not, by themselves. Therefore, the voters play an active role in the security of the e-voting scheme, and the scheme's usability becomes of crucial importance.

Vote verification typically is not present in traditional paper-based voting schemes, therefore it is likely that voters are not familiar with the tasks that are

associated with it. Furthermore, voting is no everyday activity and any election includes a share of new voters. Hence, we cannot expect that training and learning can mitigate usability issues. As a consequence, the usability of any e-voting system has to be studied thoroughly before its usage in real elections. In this paper we present and discuss methods for assessing the usability metrics of end-to-end verifiable e-voting systems via user studies. Hereby, we focus on the tasks of vote casting and verification. We discuss quantitative metrics, demographic data and the user study setting based on past usability studies that we carried out and a detailed literature review. We show that determining the effectiveness of a verification constitutes a particular challenge and has to be considered from early on when planning and designing the user study. We deliberately exclude expert evaluations, such as walkthroughs [35], because they do not entail end users as study participants.

2 End-to-End Verifiable E-Voting Schemes

E-voting schemes are based on cryptographic protocols in order to provide security properties like vote privacy or eligibility. Voters interacting with the e-voting scheme have no means to verify, that the scheme processed their votes correctly and therefore have to trust that correct processing occurs. End-to-end verifiable e-voting schemes [12] enable individual voters to verify that their votes have been processed correctly. Hereby, no trust in the voting scheme, the voters' personal computers, election officials, or external observers is required [6]. End-to-end verifiability can be subdivided[3] into the following components: (1) *cast-as-intended* meaning the cast vote corresponds to the voter's intent, (2) *recorded-as-cast* meaning the recorded vote matches the cast vote, (3) *tallied-as-recorded* meaning that all recorded votes are correctly included in the tally and (4) *eligibility* meaning that only the votes of eligible voters are tallied.

The tallied-as-recorded as well as the eligibility verifiability can be executed by observers and technically-adept users for all voters, because the information that is required to perform these verification types is publicly available, and only a certain share of voters or observers is required to do this. The recorded-as-cast verifiability has to be initiated by the voters, since solely the voters have knowledge about their participation in the election. Therefore, they play an active part in this component. The most challenging component from the voters' perspective is given by the cast-as-intended verifiability. Because the voters' intents are protected by vote privacy, only the voters themselves can perform verification[4]. Furthermore, they have to determine the outcome of the verification by themselves and act properly in case they uncover a manipulation.

[3] We use the subdivision for usability investigation purposes. Note, that the components alone do not replace end-to-end verifiability [14].

[4] A scheme for delegation has been proposed [17] whereby the complexity is shifted towards vote casting indicating that even if delegation is possible, human factors are still important.

3 The Impact of Usability

E-voting systems have been investigated in several works which confirm that the usability of the e-voting scheme is crucial. Errors rooted in a poor voting client usability can propagate to the tallying result and negatively impact the election's integrity. Therefore, poor usability renders all verification useless.

The usability of vote casting on early e-voting schemes, which are paper punch cards and lever machines, has been studied in several works [18, 23, 10, 11]. Although paper ballots are superior in terms of usability, compared to paper punch cards and lever machines and have lower error rates than punch pards, lever machines and even Direct Recording Electronics (DREs) [10]. An investigation of different DREs, which are the computers used for voting in polling stations, reveals that between 1% and 8% of cast votes do not match the voters' true intentions [13, 19] which could well flip the outcome of a first-past-the-post election.

Several verifiable e-voting schemes have been investigated in the literature. Not all of the investigated themes offer end-to-end verifiability, but some degree of verifiability. Participants in a user study of the Norwegian e-voting scheme [22] could not determine whether their votes were submitted [20]. The usability of the Benaloh Challenge was perceived as very poor [1], only between 10% and 43% of participants were able to complete verification [47, 1, 34] and the Benaloh Challenge was proven to be ineffective from a game-theory perspective [15]. A comparative usability evaluation of three Internet voting schemes by Kulyk *et al.* [28], each of which had a different level of system security, revealed that participants were willing to sacrifice 26 points on the System Usability Scale [8] when they were informed about the different degrees of system security. The first scheme was a simple click form (the least secure), the second a return code scheme, and the third used a combination of voting and return codes (being the most secure).

4 Study Design Challenges

In this section we describe study design challenges in the scope of e-voting systems, based on our study design experiences and the literature we provide guidelines to cope with these challenges. Since not all challenges can be addressed by a generic guideline, we use the challenges in the following section to derive guidelines for specific usability metrics.

4.1 Election Setting

The election setting is the specific election scenario that is provided to the participants within the study. It encompasses the election that the participant participates in, e.g., a university council or parliament election, the number of races and all aspects that concern the election and the participants' role in it.

Elections can have various stakes; governmental elections are high-stake whereas university council elections are usually low-stake. Simple polls, e.g., asking for food preferences, have an even lower stake. The stake of the election, however, can impact the participants' behaviors and the study's ecological validity, which refers to the extent to which the results of an experiment can be applied to real-world conditions [39]. Therefore, the setting can influence attitudes towards the usage of provided e-voting system features, such as verification.

Selker *et al.* [41] recommend, based on an analysis of previous voting user studies, that the study setting should be closely related to a real-world election in order to strengthen the ecological validity.

4.2 Participant Vote Privacy

Any user study collects data of the participants in order to assess the investigated scheme's usability. While some data collection methods target very specific data types (e.g., the time stamping of actions), others such as screen recording collect a plethora of data that might also be privacy-sensitive. The voting options that the study participants marks, can be part of these data and therefore the participants' vote privacy can be compromised depending on the study design and the measurements that are taken. Vote privacy, however, is a quite delicate aspect and the disclosure of voting preferences is forbidden by the law (e.g., [45]) and should therefore be also preserved in user studies. Participants are not aware of the vote privacy aspect and tend to vote for the same candidate as in a real election [46]. This introduces a trade-off between maintaining the participants' vote privacy and the measurements. The examiners have to decide whether they wish to either maintain vote privacy by measurements that do not compromise it or adjust the study design to take the compromising measures. We will emphasize on vote privacy when we discuss different measurement methods in Section 5.

4.3 Social Acceptability Bias

The *social acceptability bias* [24, 36] is the tendency of participants to give socially acceptable answers rather than answering in a way that reflects their true opinions. Therefore, participants might act differently as they would act in a real election. This introduces a challenge in designing e-voting user studies, since the social acceptability bias could impact the user study results, especially in the scope of verification. A possibility to offset the social acceptability bias is the introduction of a fictitious research goal. Budurushi *et al.* [9] used a *cover story* and told participants in the study briefing that their goal was to investigate democracy development, the general acceptability of e-voting and usability. During the debriefing the participants were told the actual research goal.

4.4 Mental Tasks

In any usability study, the researchers require knowledge about the correct execution of required tasks in order to able to measure deviations from the

correct execution. The correct execution might contain specific tasks that cannot be measured directly. Particularly challenging are tasks that the participants perform mentally. For instance, the Benaloh Challenge requires the comparison of hash values and voting options. If both match, the verified vote was cast-as-intended, therefore, a user study needs to confirm whether participants indeed perform these tasks.

4.5 Demographic Data

Demographic data is data regarding the study participants and is necessary for the determination of whether the individuals in a particular study are a representative sample of the target population for generalization purposes. Demographic data, in general, includes age, gender, occupation and education level. In e-voting specific studies the general demographic data is not sufficient, since opinions and previous voting experiences can impact the study outcome.

Furthermore, the participants' attitudes regarding the usage of e-voting in general could impact their performance and answers in the user study. Therefore, the demographics questionnaire should include questions that ask for the participants' general attitudes towards e-voting.

4.6 Motivation Interference

For the usage of any feature in any kind of system a study participant has to be motivated to do so. The construct of usability does not encompass the motivation to attempt a task in the first place. Instead of not being able to complete a task, participants might lack the motivation to attempt it. Verification in e-voting schemes is a not anticipated extra task that is not present in most traditional paper-based voting solutions. Furthermore, there are neither media campaigns that advertise verification as "positive" nor are there other incentives for the participants to verify. The investigation of usability, however, requires that the participants at least attempt to verify. Therefore, examiners need to make sure that the participants attempt verification.

5 Usability Metrics

According to ISO 9241-11 [43] the construct *usability* is defined as the *effectiveness*, *efficiency* and *satisfaction* with which specified users achieve specified goals in particular environments. ISO 9241-11 is also used by the NIST [29] for investigating voting systems. In particular, the criteria are defined as:

Effectiveness The accuracy and completeness with which specified users can achieve specified goals in particular environments.

Efficiency The resources expended in relation to the accuracy and completeness of goals achieved.

Satisfaction The comfort and acceptability of the work system to its users and other people affected by its use.

5.1 Effectiveness

Effectiveness means either the accuracy and completeness with which voters cast votes or the accuracy and completeness with which voters verify votes. Based on this definition, it is crucial to determine whether a study participant indeed cast and/or verified a vote successfully - the *binary success* [3] - or the process that participants made - the *level of success* [3].

At first the examiners need to determine the sequence of actions that is required in order to cast and/or verify a vote. The progress in this sequence has to be captured accurately in order to determine the effectiveness. Capturing the progress, however, constitutes a particular challenge when investigating end-to-end verifiable e-voting schemes for two reasons: (1) the participants' vote privacy might have to be preserved and (2) not each task within the action sequence can be assessed directly. Since the participants vote, process capturing might break their vote privacy. If vote privacy is important in the user study, proxy measurements that do not interfere with vote privacy are required. In the following we discuss several progress capturing methods as well as their relation to the challenges.

Observation The examiner could *observe* the participants to determine their progress [3]. If a real election is investigated, observation is not possible since many countries demand vote privacy in the voting booth [5]. Vote privacy is also compromized in the lab setting, because the examiner could see what the participants vote for. It furthermore introduces social acceptability bias, because the participants might alter their behavior in order to match the examiner's expectations. Finally, mental tasks cannot be assessed reliably by an examiner. Several studies address these problems by unobtrusive observation [42, 34] in which the examiner is present in the lab and administers the study, but cannot observe the participants interactions with the e-voting system while voting and verification. In case the study scenario does not demand a real vote, the participants can be provided with voting instructions. Several studies use intent cards in order to provide written voting instructions [18, 34, 28]. Participants in an e-voting study by De Jong *et al.* struggled in remembering their verbal voting instructions [26], therefore instructions should be written.

Visual Recording *Visual Recording* [3] captures the participants' screens and renders a video from it or films the participants' interactions with a camera, such that the resulting video can be analyzed after the user study. This capturing method is objective and there is no influence by the examiner's behaviour. Since the entire interaction is recorded, the voting options that the participants choose, are part of the recordings which breaks the participants' vote privacy. Therefore, the same aspects regarding vote privacy like in the observation apply here. However, if the study scenario demands that the participants cast votes that match their real intents visual recordings cannot be used at all. Furthermore, as shown by Conrad *et al.* [13] video recording might be unreliable, since 0.5% of the contests the participants voted in were not visible on the video.

Self Reporting Another possibility to access the progress is asking the participants directly whether they performed the required actions [3]. In doing so no recordings (e.g., screen recording) are required and vote privacy is maintained as long as the participants are not asked for the voting option. Self reported answers however suffer from a few drawbacks. Since self reporting is reactive, the participants have a big influence on the given answer. The participants might lie to the examiner or in the questionnaire because of the social acceptability bias [24] or because of misperceptions. For example, in a study of Helios about 25% of participants thought they successfully cast a vote although they did not [1]. 25.8% of participants in a user study by Marky *et al.* [34] stated in a questionnaire that they did verify their votes whereas in reality they failed.

Thinking Aloud In the *thinking aloud* method [7], the participants are encouraged to verbally express thoughts during the interaction with the investigated scheme. Therefore, one might assume that the participant will also comment on the tasks that they are performing. While thinking aloud can provide important insights into the participants' interactions, it is not reliable in determining effectiveness and the fulfillment of tasks, since each participant follows a different "strategy" while commenting. Furthermore, the thinking aloud method might impact the measurement of other metrics (e.g., completion time).

Eye Tracking Eye-tracking can be used to investigate the participants' gazes and their durations [3]. However, while looking at a certain display area, e.g., the one displaying a verification code, it can not assured, that the user indeed executes the required mental action, e.g., comparing the code to another. Therefore, eye tracking can only be employed in settings without mental tasks. Eye-tracking has been used in completed user studies in order to identify eye movements and gazes [27, 40].

Proxy Measures Often it is not possible to capture the participants' task success. In the e-voting setting, this might be because maintaining the vote privacy interferes with performance capturing, but also if the task is performed mentally, such as comparing data. To be able to capture the process accurately in such a situation, *proxy measures* [3] can be used. This refers to a measure that helps measuring the task, but requires some setup adjustments.

A common proxy measures when investigating e-voting schemes are deliberate *manipulations* [41]. In end-to-end verifiable e-voting schemes, the user frequently has to compare data, e.g., verification codes. To capture whether the participants have indeed compared the data, the data could be manipulated. Furthermore, the user has to be instructed on how to ask if he or she uncovers a manipulation.

The instruction of deliberate manipulations also impacts other metrics of the experiment. Participants might be less satisfied, if they experienced a manipulation. Therefore, other parameters of the usability study have to adjusted in order to account for that: (1) the participants could interact multiple times with

the system, with and without a deliberate manipulation or (2) the study is in between-subjects design, such that one group is confronted with a manipulation and the other group is not.

Error Rates Error rates are an alternative to completion rates to assess effectiveness. In the scope of e-voting systems this refers to the relationship between the voter's intention and the real outcome [13, 33]. MacNamara *et al.* [33] investigate in the usability of the DualVote VVPAT system and measured an error rate of 11.4%. However, this was not rooted in usability issues, the reason was a technical problem of the voting machine. Voter performance might be worse in real elections, since there might be pressure by other waiting voters [13].

5.2 Efficiency

According to the ISO standardization, efficiency refers to the resources expended in relation to the accuracy and completeness of goals achieved [43]. The most commonly used method in the literature to assess efficiency is the completion time, meaning the time that a participant required in order to complete all action that are required to cast or to verify a vote. Two types of completion times have been assessed in user studies in the literature: (1) the *ballot completion time* which refers to the time required by participants to mark a ballot and (2) the *verification completion time* which is the time participants required to successfully complete verification.

Ballot Completion Time The ballot completion time has been investigated by several works [1, 10, 19, 33]. The average completion time is dependent on the ballot and therefore, is identical when comparing different e-voting systems. Byrne *et al.* [10] found out that the identical ballot in different representations and voting systems requires the same completion time. In particular, they investigated arrow ballots, bubble ballots, punch cards, and lever machines which all required roughly 231 seconds for ballot completion. Everett *et al.* [19] could not find significant differences between bubble ballots, punch cards, lever machines and DREs, confirming that the completion times are ballot-dependent. Studies of Prêt à Voter implementations, which uses a very specific ballot design, show that participants require more time to mark their votes [1].

Verification Completion Time The time required for verification has been investigated in several user studies [1, 34]. Marky *et al.* [34] investigated three different interfaces of the Benaloh Challenge and found significant differences in the usage of a mobile verification device and a verification website. This shows that an interface can well impact the duration of verification.

5.3 Satisfaction

Satisfaction refers to the comfort and acceptability of the work system to its users and other people affected by its use [43]. It can be assessed by either a *standardized questionnaire* or a *non-standardized* one created by the experimenter.

System Usability Scale The System Usability Scale (SUS) [8] has been used extensively in e-voting user studies [18, 10, 21, 48, 1, 28, 34, 31, 32, 2]. Therefore, it can be used to compare the subjective usability of a "new" e-voting scheme to a range of different existing e-voting systems.

Acemyan *et al.* [1] measured an SUS score of 20.0 when investigating the usability of the Helios implementation of the Benaloh Challenge. Marky *et al.* [34] investigated the same process, but measured 76.5. The difference results from the different scopes in both studies: While Acemyan *et al.* measured the verification task only, Marky *et al.* accessed the SUS score of the voting client, including ballot marking and vote casting. Acemyan *et al.* also measured the SUS score of the voting client which resulted in a similar SUS score to the one obtained by Marky *et al.*. Therefore, it is likely that the subjective assessments from the verification significantly differs from the satisfaction related to the voting client.

The first SUS question is *I think that I would like to use this system frequently.* When conducting the user study published in [34] we found out that the answer of this question can correlate with the participants' general attitudes towards e-voting. In case participants expresses a negative attitude, their answers to this question could not be affirmative and vice versa. This might distort the SUS measurement (and possibly others), since the question aims to investigate usability and not the participants' attitudes. Therefore, it is important to access the attitude towards e-voting in the demographic questions. Since, there is no possibility to correct the participant's answer, correlations should be reported as a limitation.

User Experience Questionnaire According to ISO 9241-210 [44] *User experience* is defined as *a person's perceptions and responses that result from the use or anticipated use of a product, system or service.* The User Experience Questionnaire (UEQ) [30] can be used to measure the overall experience of the participants with the e-voting system they interact with. It covers attractiveness as well as usability aspects by measuring perspicuity, efficiency, dependability, but also the hedonic aspects of stimulation and novelty. Therefore, it broadens the investigation. The UEQ has been used in many user experiments, but so far in one voting experiment [16]. Considering the complete user experience helps to design protocols which are usable, bring satisfaction and meet the voters' expectations during the voting and verification phases.

Non-Standardized Questionnaires Voter satisfaction can furthermore be assessed by *non-standardized questionnaires* developed by the examiners depending on the study's specific purpose. Those questionnaires have been used in several

e-voting studies [33, 27, 5, 34]. It is, however, encouraged by related works [38, 10, 1] to use standardized questionnaires instead of self-developed one in order to have a comparability to previous studies.

6 Usability Study Guidelines

In this section we provide a list of guidelines to cope with the *challenges*, *metric-based* guidelines and *general* guidelines that were derived from the literature and our own experiences.

6.1 Challenge-Based Guidelines

In this section we provide guidelines for coping with e-voting study challenges of the election setting, vote privacy, social acceptability bias, mental tasks, demographic data and motivation interference.

G1 Provide a ballot with recognizable candidates from a past or upcoming election and use a setting that is close to a real-world election to strengthen ecological validity.
G2 The usage of a cover story can offset the social acceptability bias. But the cover story should be explained to the participants during the debrief.
G3 The collection of written post-test data instead of direct communication with the examiner can offset the social acceptability bias.
G4 Mental tasks should be identified by the examiners before the measurement methods are determined.
G5 Participants should be asked in the demographic questionnaire whether they have already participated in an election that matches the setting from the study to capture if participants have consistent voting experiences.
G6 Participants should be asked about their general attitude towards e-voting in the demographic questionnaire, because it could distort user study results.
G7 Written instructions to attempt verification ensure that participants try to carry out the required tasks. The instructions should not contain detailed instructions on *how* to verify a vote and solely should instruct participants to *do* it. The instructions should be tested in a pre-study to ensure their understandability.
G8 Make sure that the motivation of participants to cast or verify their votes does not impact the usability metrics.

6.2 Metric-Based Guidelines

In the following we provide guidelines regarding the assessment of the usability metrics. Hereby, we focus on process capturing for assessing effectiveness, error rates, efficiency and questionnaires.

G9 The methods of observation, visual recording and think aloud break the participants' vote privacy and should not be used if it has to be kept.

G10 The methods of observation, visual recording, self reporting, think aloud break and eye tracking should not be used for assessing the completion of mental tasks.
G11 Thinking-aloud should not be used in conjunction with completion time.
G10 Self-reporting should be used in case where no other measurements are possible, e.g., participant opinions.
G12 Self-reporting should be used as an addition to objective measurements.
G13 Written voting instructions should used to maintain vote secrecy when visual recording or observation is used.
G14 If the participants are video taped, actions of them might not be visible on the video, therefore, the video positioning should be refined in a pre-study.
G15 If vote privacy is important throughout the study, examiners should be positioned in an *unobtrusive* way, such that they cannot see the participants interactions and interfere with them.
G16 Deliberate manipulations can used as a proxy measure to capture effectiveness.
G17 Errors can be rooted in the voting system's malfunctioning. Therefore, the malfunctioning of the e-voting system has to be ruled out.

6.3 General Guidelines

Besides the specific guidelines stated above, we derived general guidelines that concern the user study overall.

G18 The participants should be informed clearly which part of the voting scheme should be considered when answering questionnaires.
G19 A plethora of baseline data for several e-voting systems is available in the literature. The usage of the same metrics provides the opportunity to compare the investigated system to those that have already been investigated. Therefore, the SUS, ballot/verification completion time and completion rates should be measured.
G20 The assessed metrics should not be limited to usability, the bigger picture of User Experience can provide valuable insights.

7 Related Work

The usability of voting systems has been investigated in many studies and some publications focus on investigation methodologies. In the following, we describe existing guidelines and recommendations for e-voting studies that are presented in related work. Our paper aims to provide guidelines related to the challenges and usability metrics while related work provides additional guidelines for the overall protocol design.

Selker *et al.* [41] analyze three studies of polling station based voting systems and provide guidelines for future work. They compare realistic and laboratory experiments for testing voting technologies. The results reveal that real-world

tests are more valuable as they add a considerable workload to the process, that uncovers additional issues related to the environment and to the procedure in polling stations (distraction, confusion, importance of poll workers assistance). However, the collected data is consistent in both test environments and the aspects of a real-world setting do not influence ballot understanding or verification results. Providing a voting card to the participants in a mock election where they know the candidates is not efficient: participants could vote for a different candidate despite the given instructions, which impacts error tracking. For methodologies, authors recommend testing as far as possible with a real-world setting, to improve ecological validity.

Olembo and Volkamer [38] conduct a literature review focusing on usability studies for e-voting systems. They review different designs and methodologies, and provide a list of recommendations for user studies. They stress that expert evaluations are faster and more cost-effective and bring more data compared to users studies. They describe a methodology for future tests: at first an iteration of the tested protocol must be done with HCI experts and users must be involved in a second iteration with a pilot study. The final design must be re-tested with users and fields studies can be considered.

Herrnson *et al.* [25] study the importance of usability by investigating six voting schemes that are already in use, with different voting procedures, with four research methodologies. In their recommendation, they do not discuss the different aspects related to the methodology in use, but focus on usability findings and impacts for next studies. They notice that demography impacts the voters' needs, in particular the need of assistance was at least 18%. Therefore, accessibility must be taken into consideration and a proper assistance should be available.

Taha Ali and Murray [4] discuss the impact of usability on voting systems, saying that "ensuring system usability is also complicated by the fact that elections occur only rarely and voters must be able to vote with near 100% success while having little or no experience or training on that voting system". This lack of experience can be extended to poll workers and election officials, and concerns are on whether voters will be able to cast a vote successfully, but also able to go through the verification process. Therefore, the authors state that a trainee at an early stage for voters and poll workers must be done to increase usability results.

8 Conclusion

The usability of end-to-end verifiable e-voting schemes is equally important as their security and directly impacts it. Therefore, human factors should be considered from early on when designing usable end-to-end verifiable e-voting systems that are intended to be put into practice. In this paper we provide guidelines for investigating vote casting and vote verification in e-voting schemes in user studies. The guidelines and remarks are pulled from the literature and based on the authors' experiences and aim to inform future user studies of e-voting schemes. In this paper we focus on quantitative metrics as well as demographics

and the study setting. It would be beneficial to see guidelines for qualitative research as a part of future work.

Acknowledgements

This work has been co-funded by the DFG as part of project "Area D.1" within the RTG 2050 "Privacy and Trust for Mobile Users" and by the Horst Görtz Foundation. We acknowledge support from the Luxembourg National Research Fund (FNR) for funding, in particular Marie-Laure Zollinger was supported by the FNR-INTER-VoteVerif project and FNR-INTER-SeVoTe project.

References

1. Acemyan, C.Z., Kortum, P., Byrne, M.D., Wallach, D.S.: Usability of voter verifiable, end-to-end voting systems: Baseline data for helios, prêt à voter, and scantegrity ii. USENIX Journal of Election Technology and Systems (JETS) **2**(3), 26–56 (2014)
2. Acemyan, C.Z., Kortum, P., Byrne, M.D., Wallach, D.S.: Summative usability assessments of star-vote: A cryptographically secure e2e voting system that has been empirically proven to be easy to use. Human factors pp. 1–24 (2018)
3. Albert, W., Tullis, T.: Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes (2013)
4. Ali, S.T., Murray, J.: An overview of end-to-end verifiable voting systems. In: Real-world electronic voting: Design, analysis and deployment. pp. 171–218. CRC Press (2016)
5. Bederson, B.B., Lee, B., Sherman, R.M., Herrnson, P.S., Niemi, R.G.: Electronic voting system usability issues. In: SIGCHI Conference on Human Factors in Computing Systems (CHI). pp. 145–152. ACM (2003)
6. Benaloh, J., Rivest, R., Ryan, P.Y., Stark, P., Teague, V., Vora, P.: End-to-end verifiability pp. 1–7 (2015), https://arxiv.org/pdf/1504.03778.pdf
7. Boren, T., Ramey, J.: Thinking aloud: Reconciling theory and practice. IEEE transactions on professional communication **43**(3), 261–278 (2000)
8. Brooke, J.: SUS - a quick and dirty usability scale. Usability Evaluation in Industry **189**(194), 4–7 (1996)
9. Budurushi, J., Renaud, K., Volkamer, M., Woide, M.: An investigation into the usability of electronic voting systems for complex elections. Annals of Telecommunications **71**(7-8), 309–322 (2016)
10. Byrne, M.D., Greene, K.K., Everett, S.P.: Usability of voting systems: Baseline data for paper, punch cards, and lever machines. In: SIGCHI Conference on Human Factors in Computing Systems (CHI). pp. 171–180. ACM (2007)
11. Campbell, B.A., Byrne, M.D.: Now do voters notice review screen anomalies? a look at voting system usability. In: Conference on Electronic Voting Technology/Workshop on Trustworthy Elections (EVT/WOTE). USENIX Association (2009)
12. Chaum, D.: Secret-ballot receipts: True voter-verifiable elections. IEEE Security & Privacy **2**(1), 38–47 (2004)
13. Conrad, F.G., Bederson, B.B., Lewis, B., Peytcheva, E., Traugott, M.W., Hanmer, M.J., Herrnson, P.S., Niemi, R.G.: Electronic Voting Eliminates Hanging Chads but Introduces New Usability Challenges. International Journal of Human-Computer Studies **67**(1), 111–124 (2009)

14. Cortier, V., Galindo, D., Küsters, R., Mueller, J., Truderung, T.: Sok: Verifiability notions for e-voting protocols. In: Symposium on Security and Privacy (S&P). pp. 779–798. IEEE (2016)
15. Culnane, C., Teague, V.: Strategies for voter-initiated election audits. In: International Conference on Decision and Game Theory for Security (GameSec). pp. 235–247. Springer (2016)
16. Distler, V., Zollinger, M.L., Lallemand, C., Rønne, P.B., Ryan, P.Y., Koenig, V.: Security–visible, yet unseen? how displaying security mechanisms impacts user experience and perceived security [to appear]. In: CHI Conference on Human Factors in Computing Systems. CHI '19, ACM (2019)
17. Escala, A., Guasch, S., Herranz, J., Morillo, P.: Universal cast-as-intended verifiability. In: International Conference on Financial Cryptography and Data Security (FC). pp. 233–250. Springer (2016)
18. Everett, S.P., Byrne, M.D., Greene, K.K.: Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **50**(24), 2547–2551 (2006)
19. Everett, S.P., Greene, K.K., Byrne, M.D., Wallach, D.S., Derr, K., Sandler, D., Torous, T.: Electronic voting machines versus traditional methods: Improved preference, similar performance. In: SIGCHI Conference on Human Factors in Computing Systems (CHI). pp. 883–892. ACM (2008)
20. Fuglerud, K.S., Røssvoll, T.H.: An evaluation of web-based voting usability and accessibility. Universal Access in the Information Society **11**(4), 359–373 (2012)
21. Gibson, J.P., MacNamara, D., Oakley, K.: Just like paper and the 3-colour protocol: A voting interface requirements engineering case study. In: International Workshop on Requirements Engineering for Electronic Voting Systems. pp. 66–75. IEEE (2011)
22. Gjøsteen, K.: The Norwegian internet voting protocol. In: International Conference on E-Voting and Identity (VoteID). pp. 1–18. Springer-Verlag (2011)
23. Greene, K.K., Byrne, M.D., Everett, S.P.: A comparison of usability between voting methods. In: Electronic Voting Technology Workshop (EVT). USENIX Association (2006)
24. Grimm, P.: Social desirability bias. Wiley International Encyclopedia of Marketing (2010)
25. Herrnson, P.S., Niemi, R.G., Hanmer, M.J., Bederson, B.B., Conrad, F.G., Traugott, M.: The importance of usability testing of voting systems. In: Electronic Voting Technology Workshop (EVT) (2006)
26. de Jong, M., van Hoof, J., Gosselt, J.: User research of a voting machine: Preliminary findings and experiences. Journal of Usability Studies **2**(4), 180–189 (Aug 2007)
27. Karayumak, F., Kauer, M., Olembo, M.M., Volk, T., Volkamer, M.: User study of the improved helios voting system interfaces. In: Workshop on Socio-Technical Aspects in Security and Trust (STAST). pp. 37–44. IEEE (2011)
28. Kulyk, O., Neumann, S., Budurushi, J., Volkamer, M.: Nothing comes for free: How much usability can you sacrifice for security? IEEE Security & Privacy **15**(3), 24–29 (2017)
29. Laskowski, S.J., Autry, M., Cugini, J., Killam, W., Yen, J.: Improving the usability and accessibility of voting systems and products. NIST Special Publication **500**, 256 (2004)
30. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: HCI and Usability for Education and Work (2008)

31. Mac Namara, D., Gibson, P., Oakley, K.: A preliminary study on a dualvote and prêt à voter hybrid system. In: Conference for E-Democracy and Open Government. p. 77. Edition-Donau-Univ. Krems (2012)
32. Mac Namara, D., Scully, T., Gibson, P.: Dualvote addressing usability and verifiability issues in electronic voting systems (2011), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.399.7284
33. MacNamara, D., Carmody, F., Scully, T., Oakley, K., Quane, E., Gibson, J.P.: Dual vote: A novel user interface for e-voting systems. In: International Conference on Interfaces and Human Computer Interaction. pp. 129–138. IADIS (2010)
34. Marky, K., Kulyk, O., Renaud, K., Volkamer, M.: What did i really vote for? In: CHI Conference on Human Factors in Computing Systems (CHI). pp. 176:1–176:13. ACM (2018)
35. Marky, K., Kulyk, O., Volkamer, M.: Comparative usability evaluation of cast-as-intended verification approaches in internet voting. In: "SICHERHEIT 2018". pp. 197–208. Lecture Notes in Informatics (LNI), Gesellschaft für Informatik (2018)
36. Nancarrow, C., Brace, I.: Saying the "right thing": Coping with social desirability bias in marketing research. Bristol Business School Teaching and Research Review **3**(11), 1–11 (2000)
37. Neumann, S.: Evaluation and Improvement of Internet Voting Schemes Based on Legally-Founded Security Requirements. Ph.D. thesis, Technische Universität Darmstadt (2016)
38. Olembo, M.M., Volkamer, M.: E-voting system usability: Lessons for interface design, user studies, and usability criteria. In: Human-Centered System Design for Electronic Governance, pp. 172–201. IGI Global (2013)
39. Patrick, A.: Ecological validity in studies of security and human behaviour. In: Symposium on Usable Privacy and Security (SOUPS) (2009)
40. Realpe-Muñoz, P., Collazos, C.A., Hurtado, J., Granollers, T., Muñoz-Arteaga, J., Velasco-Medina, J.: Eye tracking-based behavioral study of users using e-voting systems. Computer Standards & Interfaces **55**, 182–195 (2017)
41. Selker, T., Rosenzweig, E., Pandolfo, A.: A methodology for testing voting systems. Journal of Usability Studies **2**(1), 7–21 (Nov 2006)
42. Sherman, A.T., Carback, R., Chaum, D., Clark, J., Essex, A., Herrnson, P.S., Mayberry, T., Popoveniuc, S., Rivest, R.L., Shen, E., et al.: Scantegrity mock election at takoma park. In: International Conference on Electronic Voting (EVOTE). pp. 45–61. LNI, Gesellschaft für Informatik (2010)
43. Standardization, I.O.F.: Iso 9241-11: Ergonomics of human system interaction – part 11: Guidance on usability (1998)
44. Standardization, I.O.F.: Iso 9241-210: Part 210: Human-centred design for interactive systems (2015)
45. Strafgesetzbuch (StGB): §107c Verletzung des Wahlgeheimnisses. https://www.gesetze-im-internet.de/stgb/__107c.html
46. Van Hoof, J.J., Gosselt, J.F., de Jong, M.D.: The reliability and usability of the nedap voting machine: A pilot study. University of Twente Faculty of Behavioural Sciences Department of Technical and Professional Communication (2007)
47. Weber, J.L., Hengartner, U.: Usability study of the open audit voting system helios. http://www.jannaweber.com/wp-content/uploads/2009/09/858Helios.pdf (2009), [Online; accessed: 12-June-2018]
48. Winckler, M., Bernhaupt, R., Palanque, P., Lundin, D., Leach, K., Ryan, P., Alberdi, E., Strigini, L.: Assessing the usability of open verifiable e-voting systems: A trial with the system prêt à voter. In: International Conference on Theory and Practice of Electronic Governance (ICEGOV). pp. 281–296. ACM (2009)