Noname manuscript No.
(will be inserted by the editor)

An Investigation into the Usability of Electronic Voting Systems for Complex Elections

Jurlind Budurushi · Karen Renaud
Melanie Volkamer · Marcel Woide

Received: date / Accepted: date

Abstract Many studies on electronic voting evaluate their usability in the context of simple elections. Complex elections, which take place in many European countries, also merit attention. The complexity of the voting process, as well as that of the tallying and verification of the ballots, makes usability even more crucial in this context. Complex elections, both paper-based and electronic, challenge voters and electoral officials to an unusual extent. In this work we present two studies of an electronic voting system that is tailored to the needs of complex elections. In the first study we evaluate the effectiveness of the ballot design with respect to motivating voters to verify their ballot. Furthermore, we identify factors that motivate voters to verify, or not to verify, their ballot. The second study also addresses the effectiveness of the ballot design in terms of verification, but this time from the electoral officials' perspective. Last, but not least, we evaluate the usability of the implemented EasyVote prototype from both the voter and electoral official perspectives. In both studies we were able to improve effectiveness, without impacting effi-

This project (HA project no. 435/14-25) is funded in the framework of Hessen ModellProjekte, financed with funds of LOEWE Landes-Offensive zur Entwicklung Wissenschaftlich-konomischer Exzellenz, Frderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

Jurlind Budurushi
Technische Universtität Darmstadt, Germany
E-mail: jurlind.budurushi@secuso.org

Karen Renaud
University of Glasgow, UK
E-mail: karen.renaud@glasgow.ac.uk

Melanie Volkamer
Technische Universtität Darmstadt, Germany
E-mail: melanie.volkamer@secuso.org

Marcel Woide
Technische Universtität Darmstadt, Germany
E-mail: marcel.woide@secuso.org

ciency and satisfaction. Despite these usability improvements it became clear that voters who trusted the electronic system were unlikely to verify their ballots. Moreover, these voters failed to detect the "fraudulent" manipulations. It is clear that well-formulated interventions are required in order to encourage verification and to improve detection of errors or fraudulent attempts.

Keywords Electronic Voting · Usability · Verification · Paper Audit Trails · Complex Elections

1 Introduction

The idea of electronic voting machines is almost as old as the electricity that powers them. One of the earliest efforts to develop electronic voting machines dates back to 1875 with the model developed by Martin de Brettes [5]. Despite public interest, none of the early electronic voting machines was ever actually used [22]. 100 years later, in the 1970s, electronic voting machines were deployed in U.S elections [43] and the Netherlands [27], and latterly in Germany in the late 1990s [1].[1]

The *first* generation electronic voting machines[2], such as those developed by Avante, Diebold, ES&S, Sequoia and Nedap, and those from the academic research [7, 12, 35, 36, 49], were black-box systems. Voters cast their votes using an electronic voting machine. The machine simply stored the votes to be tallied once the election was over. Election fraud was impossible to detect if a voting machine were compromised. Many countries, including Germany and the Netherlands, discontinued their use of these machines.

The *second* generation electronic voting machines, such as those developed by Smartmatic and those from the academic research [3, 4, 13, 48], provide voters with an independent physical proof of their vote. Usually this is provided in the form of a so-called paper audit trail[3]. Unlike the *second* generation electronic voting machines, the *third* generation electronic voting machines [44, 45, 47] do not store the votes electronically and thus the election result is based only on paper audit trails. These paper audit trails consist of a human- and machine-readable parts (e.g. a QR-Code or a RFID-Chip). These can be tallied either manually or electronically. Hence, the *second* and *third* generation electronic voting machines, in contrast to the *first* generation, facilitate fraud detection. This can be achieved by performing random audits [25, 26, 40–42]. It can also be achieved by carrying out a full manual tally, something that can be semi-automated in case of *third* generation electronic voting machines.

[1] For a worldwide *status quo* of electronic voting visit the link: `http://www.e-voting.cc/wp-content/uploads/2012/03/e-voting_worldmap_2015.pdf`, last accessed 16 November 2015.

[2] For a more detailed classification of electronic voting systems refer to [15, 18], and for voting forms to [23].

[3] The idea of paper audit trails was invented by Mercuri [30], and later associated by Rivest and Wack with the notion of software-independence [34].

In this work we focus on the *third* generation electronic voting machines, as used in Argentina [32] and Ecuador [33]. Their use is being considered in the Netherlands [27] as proposed by [45]. In order to ensure the integrity of the election result, these machines rely on the following assumptions being true: (1) Voters will verify that the human-readable part of their paper audit trail reflects their intentions; (2) Electoral officials will verify that the human-readable part matches the machine-readable part.

Unfortunately, a number of studies [2, 14, 19, 38] showed that voters were unlikely to verify their paper audit trails. This is curious because such study participants often consider paper audit trails to be important [29]. Even electoral officials might not verify as effectively as they ought to. The two assumptions become even more unrealistic in the context of complex elections, i.e. elections with complex ballots and voting rules. Hence, the goal of this work was to test the validity of these assumptions in the context of complex elections.

In order to achieve our goal we focus on the local elections in Hesse (as an exemplar of a complex election) and on the EasyVote voting scheme [45], based on the findings presented in [10] and the fact that we could use and adapt the open source EasyVote prototype[4]. The contribution of this work can be summarised as follows:

1. For complex elections, we carry out an initial investigation into:
   - the validity of the assumptions underlying the use of *third generation* electronic voting machines, both from the voter and the electoral official perspectives.
   - the usability[5] of a *third generation* electronic voting machine prototype in accordance with the ISO 9421-11 standard [20].
   - voter perceptions with respect to the verification of the human-readable part of the paper audit trail.
   - factors that motivate voters to verify, or not to verify, the paper audit trail.
2. We propose a research design for testing hypotheses related to ballot verification by voters and electoral officials.

A subsequent investigation with a larger number of participants is required to follow up our findings, in order to confirm our results.

2 Background

In this Section we first describe the local elections in Hesse. We then briefly introduce the EasyVote voting scheme [45] and describe the components of the EasyVote ballot. Lastly we introduce the definition of usability according to the ISO 9421-11 standard [20].

[4] `https://github.com/SecUSo/EasyVote`, last accessed 16 November 2015.

[5] According to Volkamer *et al.* [46] evaluating the usability of electronic voting systems is important and critical for trust establishment.

2.1 Local elections in Hesse

In this Section we describe the vote casting and tallying process in Hesse local elections.

2.1.1 Vote casting process

The vote casting process in the Hesse local elections is similar to elections in Luxembourg, Belgium and other German states (Bavaria, Bremen, Hamburg), in that it shares similar voting rules. In the following text we introduce the concrete Hesse voting rules for local elections by using the city of Darmstadt as an example:

- The voter can cast 71 direct votes.[6]
- The voter can assign up to three votes to each candidate (cumulative voting).
- The voter can cast votes to candidates of different parties (vote splitting).
- The voter can select only a party.
- The voter can cross out candidates.

If the voter selects a party, the votes that are not directly cast are automatically assigned to the candidates of the selected party according to the list order[7]. By being able to cross out candidates the voter can influence the automatic vote assignments when a party is selected. Depending on the size of the district more than ten parties and more than 450 candidates participate in elections, which results in huge ballots, shown in Figure 1. Since this undoubtedly leads to error, so-called *healing rules* are applied during the tallying process:

- If the voter casts more than 71 direct votes for candidates of one party, only 71 votes are tallied.
- If the voter selects more than one party, and casts fewer than, or equal to, 71 direct votes, only the direct votes are tallied.

These healing rules aim to validate a ballot which might be interpreted as invalid according to the strict letter of the rules.

2.1.2 Tallying process

The tallying process consists of two phases carried out by electoral officials and monitored by an electoral officer. The first phase of the tallying process starts at the end of the election day. Electoral officials are required to perform the following steps:

[6] Note that the maximal number of votes a voter can cast, is determined by the number of seats in the parliament.

[7] Note that the automatic assignments of votes takes place in the tallying process.

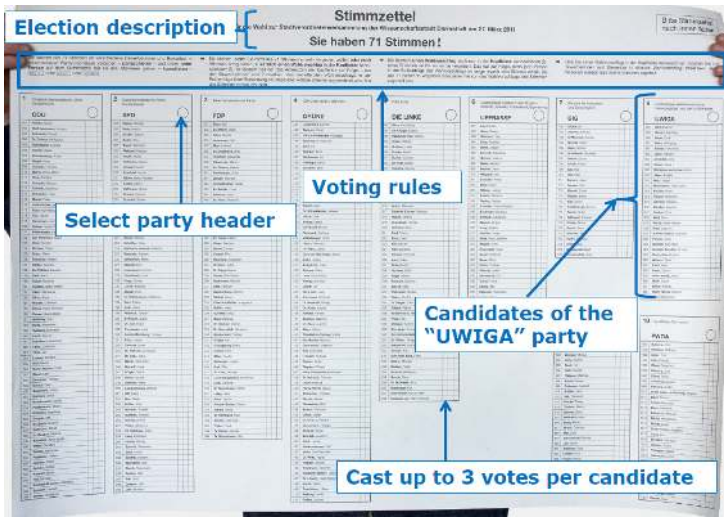

Fig. 1: Ballot of the local elections in Darmstadt 2011 (Size: 27" x 35").

1. Open the ballot boxes and ensure that the number of cast ballots matches the number of voters marked off in the electoral register.
2. Classify ballots into four *categories*: (1) Only a party is selected; (2) Candidates are directly selected (and a party is selected); (3) Invalid; (4) Cannot be assigned to any of the other categories.
3. Verify that ballots are correctly classified.
4. Manually tally the ballots of the 1st category according to the selected party.
5. Review ballots of the 4th category and assign them to the 1st, 2nd or 3rd category.
6. Manually tally the intermediate result, by considering only the ballots of the 1st and 3rd category.

The second phase takes place the day after the election. During this phase the ballots of the second category are tallied, and electoral officials are supported by tallying software, namely the PC-Wahl[8]. Electoral officials have to perform the following steps:

1. Enter the intermediate result from the first phase into the PC-Wahl.
2. Enter the first five ballots into the PC-Wahl.
3. Manually tally the first five ballots.
4. Verify that the outcome of the second and third step matches.[9]
5. Enter the rest of the ballots into the PC-Wahl.
6. Compute the final election result with PC-Wahl.

[8] `http://www.pcwahl.de/`, last accessed 16 November 2015.

[9] The PC-Wahl software is assumed to be trustworthy. Thus, this step serves as a self-control for the electoral officials.

7. Sign the printed disposition.

The process of entering ballots into PC-Wahl is performed by three electoral officials. The first electoral official narrates the voter's selection(s) on the ballot, the second enters them into the PC-Wahl, and the third verifies the correctness the process. Figure 2 shows the PC-Wahl interface for entering/recording the ballots.

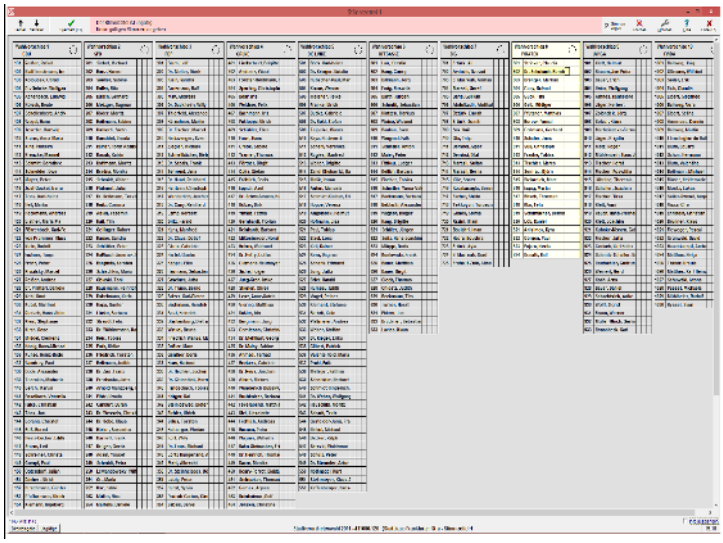

Fig. 2: The interface of PC-Wahl to enter/record the ballots.

Electoral officials who participate in the second phase are employees of the corresponding municipality and are therefore considered to have sufficient technical expertise. They participate in a introductory workshop to familiarise them with the PC-Wahl software. During the workshop, which usually lasts 30 minutes, the electoral officials have the opportunity to try out the PC-Wahl software.

2.2 The EasyVote voting scheme

In this Section we provide an overview of the EasyVote voting scheme which was first proposed by Volkamer *et al.* [45].

2.2.1 Vote casting process

During the vote casting process the voter identifies him or herself to the electoral officials, as is customary for traditional (paper-based) elections. The voter then enters the voting booth to use the electronic voting machine. The voter

prepares the ballot by selecting the preferred candidates on the voting machine. The voting machine supports and provides feedback about the current state of the ballot, specifically whether it is valid or not. The ballot is printed when the voter confirms his/her final selections. Electronic data are not retained on the machine after printing. The printed ballot consists of two parts that represent the cast votes: human- and machine-readable parts. The voter verifies that the human-readable part reflects his/her selected candidates then folds the printed ballot, exits the voting booth and deposits the ballot into the sealed ballot box.

2.2.2 Tallying process

During the tallying process electoral officials ensure that the number of cast ballots is equal to the number of voters. Afterwards electoral officials start tallying/scanning each individual ballot. Electoral officials scan the QR-Code and ensure that its content, as displayed on the monitor, matches the human-readable part of the ballot. After electoral officials have verified the content, and confirmed its correctness, the scanned ballot is added to the intermediate result, shown on a second monitor. Electoral officials, and the general public, can verify that the ballot is added correctly to the intermediate result. They repeat this process for all ballots.

2.3 EasyVote ballot

Figure 3 introduces the components of the EasyVote ballot, as proposed by Budurushi *et al.* [9].

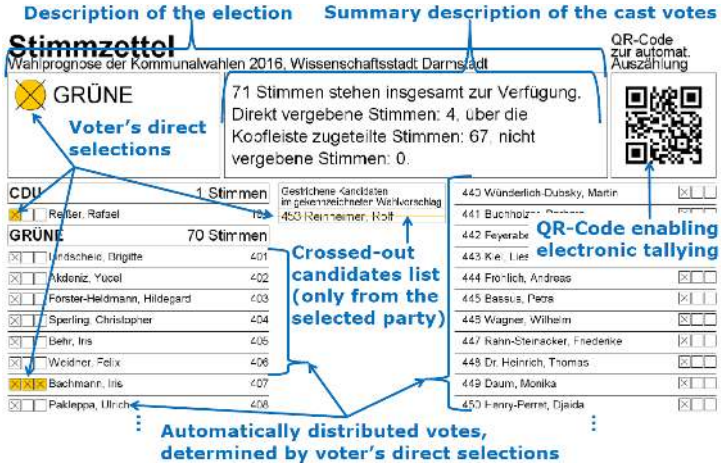

Fig. 3: The components of the EasyVote ballot.

2.4 Usability

In order to evaluate the usability of the EasyVote prototype, specifically the vote casting and tallying components, we used the ISO 9421-11 standard [20]. According to ISO 9421-11 usability comprises the following three components:

- **Effectiveness** (the ability of users to complete their task).
- **Efficiency** (the extent to which users consume resource to perform their task).
- **Satisfaction** (the level of satisfaction users experience in performing their task).

These components are defined with respect to the context of use.

3 Preliminary considerations

In this Section we introduce some preliminary considerations that are shared by both studies presented in this work.

3.1 Communicated research goal

The main challenge in the area of usable security is to avoid the potential social desirability bias [24], otherwise participants may act in a manner perceived to be socially desirable rather than acting as they actually would in a voting situation [39]. To offset social desirability a fictitious research goal was communicated instead of the actual research goal. Hence, participants of both studies were told that goal was to evaluate the usability of, and actively feed into, the implemented prototype by following the human-centred design principles for interactive systems [20, 21].

3.2 Participants' experience with electronic voting

Participants in both studies were used to the paper-based voting system, because the use of electronic voting in Germany has been discontinued since the introduction of the principle of *public nature of elections* by the German Federal Constitutional Court in 2009 [17].

3.3 Ethical considerations

An ethics commission at the authors' university lays down ethical requirements for research involving humans[10]. These requirements were met; all participants

[10] `http://www.intern.tu-darmstadt.de/gremien/ethikkommisson/index.en.jsp`, last accessed 16 November 2015.

in both studies read and signed the declaration of consent for participating in the study. Furthermore, at the end of the study, participants were debriefed regarding the actual research goal. Note that participants were able to interrupt and leave the study at any time without needing to provide a reason.

4 Study 1: Evaluation of the vote casting process

The primary goal of the study was to determine whether participants would deliberately detect manipulations of their printed ballot. We intentionally and randomly manipulated participants' ballots. We also evaluated the usability of the implemented prototype according to the ISO 9421-11 standard [20].

4.1 Methodology

4.1.1 Cover story

In order to conceal our primary goal, in addition to the communicated research goal introduced in Section 3.1, we used a cover story. Participants were told that they were taking part in a research study that emerged from a collaboration between the computer science and political science departments of the *Technische Universität Darmstadt*. The stated research focus was said to be twofold: firstly political and secondly human-centred usability focused, as follows. (1) The political science department is interested in the development of democracy in Germany. Hence, the research study aimed to compare citizens' interest and motivation to actively participate in politics. The study claimed to conduct a comparison between citizens involved in the 1960s protest movements and the citizens of today. To achieve this participants are asked to cast a vote in an election forecast of forthcoming local elections. Furthermore, to strengthen the credibility of this research focus, we used real candidate names and parties. (2) The computer science department is interested in the general acceptability of electronic voting and in evaluating the usability of the implemented EasyVote prototype.

4.1.2 Introducing manipulations: Altering the printed ballot

The manipulations used in the present study are:

1. Exchange a party, including its candidates.
2. Exchange a crossed-out candidate.
3. Exchange a directly selected candidate.
4. Exchange votes between directly selected candidates.

This set includes all manipulations used in similar previous studies [16, 38]. We intentionally and randomly manipulated the ballots by introducing one of the manipulations presented above. However, in order to ensure that

manipulations were non-trivial to detect, we introduced manipulations in a pseudo-random fashion based on the participant's direct selection(s) as follows, based on the following dependencies:

1. The participant selects only a party, implying that only the manipulation *Exchange a party, including its candidates* can be introduced.
2. The participant selects only a party and crosses out at least one candidate, implying that only the manipulation *Exchange a crossed out candidate* is randomly introduced.
3. The participant selects only one direct candidate (in combination with selecting a party), implying that only the manipulation *Exchange a directly selected candidate* is randomly introduced.
4. The participant selects one direct candidate, a party and crosses out at least one candidate of the selected party, implying that either *Exchange a crossed out candidate* or *Exchange a directly selected candidate* is randomly introduced.
5. The participant selects direct candidates and assigns them the same number of votes (in combination with selecting a party), implying that only the manipulation *Exchange a directly selected candidate* is randomly introduced.
6. The participant selects direct candidates, assigns them the same number of votes, and selects a party and crosses out at least one candidate of the selected party, implying that either *Exchange a crossed out candidate* or *Exchange a directly selected candidate* is randomly introduced.
7. The participant selects direct candidates and assigns them a different number of votes (in combination with selecting a party), implying that either *Exchange a directly selected candidate* or *Exchange votes between directly selected candidates* is randomly introduced.
8. The participant selects direct candidates, assigns them a different number of votes, and selects a party and crosses out at least one candidate of the selected party, implying that either *Exchange a crossed out candidate* or *Exchange a directly selected candidate* or *Exchange votes between directly selected candidates* is randomly introduced.

Note that the second and fourth manipulation change only the distribution of cast votes. The first and third manipulation introduce new content, e.g. a new party and/or new candidate(s), into the printed ballot.

4.1.3 Questionnaires

Pre-questionnaire. This questionnaire helps us to conceal our primary goal by assessing participants' general interest in politics.

Post-questionnaire. This questionnaire serves to compare the results of our study with the results presented by Budurushi *et al.* [11]. Thus, we compare participants' self-reported answers and their actual behaviour with respect to verifying their printed ballot. To assess participants' perceptions regarding the verification of the printed ballot, we extended this questionnaire with the

following two statements: (1) *I perceived verification of the printed ballot to be a demanding task*; and (2) *I perceived verification of the printed ballot to be an error prone task*. Furthermore, in order to identify the participants' motivations, either to verify, or not to verify, the printed ballot, we extended the questionnaire with the following open question: *What motivated you to (or not to) verify the printed ballot?*.

4.1.4 Evaluation

To evaluate the usability of the implemented vote casting prototype we defined the usability components introduced in Section 2.4, according to our context as follows:

Effectiveness: *The ability of participants to cast their vote, i.e. to cast a printed ballot that reflects their intention.* The vote-casting task included the following sub-tasks, requiring the participant to:

1. commence the vote casting process.
2. makes a selection(s).
3. print the ballot.
4. verify that the printed ballot reflects his/her intention, i.e. the participant detects the manipulations if they are present.

We measured effectiveness by observing the vote-casting process and filling in a check-list, with a box for each of these sub-tasks. Note that while measuring effectiveness we implicitly evaluated our primary goal: if participants detected that their printed ballot has been manipulated.

Efficiency: *The time spent by participants to cast their vote.* The following times are measured:

1. Time spent to read the voting rules.
2. Time spent to make selection(s).
3. Time spent to verify the printed ballot.

To measure the *Time spent to read the voting rules* we used a stopwatch and observed the enabling device. More specifically we measured the time between two events: (1) Participant enters the voting booth; and (2) The red LED of the enabling device light turning on.

Further, the implemented prototype software measured the *time spent to make selection(s)*, i.e. the time interval between starting to make selection(s) and starting the printing process. In order to measure the *Time spent to verify the printed ballot* we first measured the total time a participant spent in the voting booth and the average time the prototype software takes to print the ballot. Afterwards, we subtracted the *time spent to read the voting rules*, *time spent to make selection(s)* and the average time the prototype takes to print the ballot from the total time that the participant spent in the voting booth.

Satisfaction: *The level of satisfaction participants experience in casting their vote.* In order to measure satisfaction we used a German translation of the original SUS questionnaire [6], as improved by [28].

Finally, we conducted a qualitative analysis of the participants' responses, provided in response to the questionnaire's open questions. An open coding approach was applied. The authors independently reviewed these responses and came up with a list of codes (i.e. reasons that motivate participants to verify or to not verify their printed ballot). We considered each code, even if several codes were mentioned by a single participant. The identified codes were discussed and agreed upon by the authors.

4.2 Experimental setup, design and procedure

The experiments were carried out in our lab. The participants were randomly assigned to two groups: a control and a treatment group.

The treatment group was provided with the stimulus proposed by Budurushi *et al.* [11]: pre-printed verification instructions on the reverse of the printed ballot. The control group was provided with a printed ballot with nothing printed on the reverse. Note that we adapted the verification instructions due to the different election context addressed in the present study and the one addressed in [11]. More specifically, we replaced the upper picture and extended the first sentence with an orange highlighted and underlined adjective, namely "orange marked". Figure 4 shows the adapted instructions.

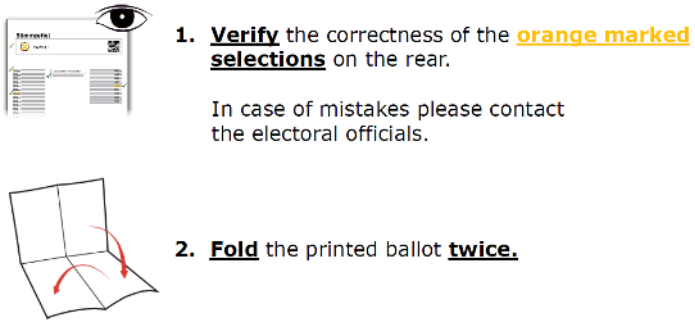

Fig. 4: The adapted verification instructions (English translation).

In the experiment each participant had to perform the following steps:

1. Read and sign the declaration of consent for participating in the study.
2. Read the cover story.
3. Fill in the demographics.
4. Fill in the questionnaire regarding the political interest.
5. Read the poster describing the vote casting process, in order to become familiar with it.[11]

[11] Note that if participants had questions regarding the vote casting process while reading the poster, they could ask the experimenter.

6. Cast a vote[12] and deposit the printed ballot into the appropriate ballot box.
7. Fill in the post-questionnaire.
8. Fill in the SUS questionnaire.
9. Debrief: Reveal the actual research goal.

If participants were not able to summarise the vote casting process correctly, the experimenter described the correct vote casting process by using the poster. If a participant detected a manipulation, debriefing took place before filling in the post-questionnaire. Furthermore, in order to ensure vote secrecy, the experimenter instructed all participants that detected the manipulation to fold and deposit their ballots into the ballot box.

4.3 Recruiting and sampling

Participants were recruited via flyers and by personal contact. 44 subjects participated (14 female, 30 male), ranging in age from 20-75, with an average age of 31.45 years ($\sigma = 13.44$) and a median age of 25. One participant had completed an apprenticeship; 20 participants had a high school degree or less; and the rest of the participants had a bachelor's degree or equivalent.

All participants were naïve with respect to the purpose of the study. In order to encourage participation we provided €10 per participant.

4.4 Materials

The materials used in the study were:

- A voting terminal, consisting of a touchscreen monitor, a computer, an enabling device and a printer.
- A voting booth.
- A poster that describes the vote casting process hanging on the outside of the voting booth.
- A shredder.
- The implemented vote casting prototype.
- Two ballot boxes, to separate ballots into two categories: (1) *Manipulation detected*; and (2) *Manipulation not detected*.[13]
- Ballots.

[12] Participants in the voting booth could print only one ballot, as proposed by Budurushi *et al.* [8]. If the participant wanted to cast a different vote from the one already printed, he/she was required to leave the voting booth, privately destroy the printed ballot, and re-enter the voting booth.

[13] Note that in the study the ballot boxes were not labelled and one of the ballot boxes was hidden in order not to confuse or bias the participants.

4.5 Results

The findings are reported in terms of *Usability evaluation* and *Participants' perception towards verification.*

4.5.1 Usability

Effectiveness. Figure 5 presents the percentage of participants that were able to cast a vote successfully, i.e. to cast a printed ballot that reflects their intention, according to the sub-tasks defined in Section 4.1.4.

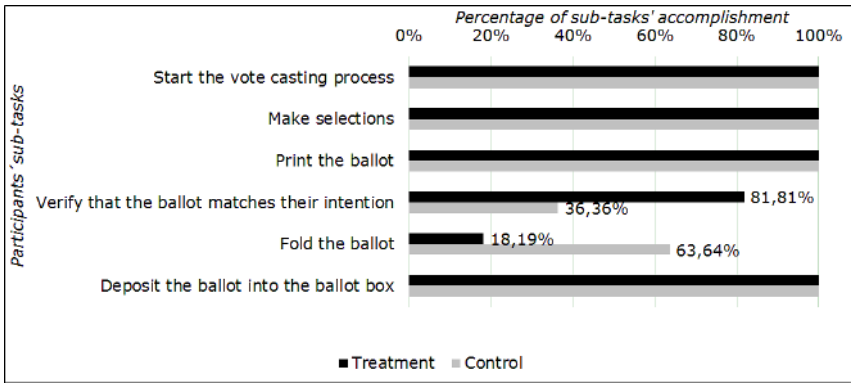

Fig. 5: The percentage of successful participants able to cast a printed ballot that reflects their intention.

The participants in the treatment group detected the randomly introduced manipulations more frequently with 81.81% (18 out of 22) compared to 36.36% (8 out of 22) of participants in the control group. In order to determine if there is a relationship between the *stimulus* and detecting manipulations between both groups, we ran a Chi-square test. The Chi-square test suggested that there is a significant difference in detecting the manipulation between the treatment and the control group ($\chi^2 = (1, N = 44) = 9.402$, $p < 0.002$).

Furthermore, Figure 6 provides an overview of the frequency of occurrence for each manipulation category, and whether a manipulation was detected or not. The manipulation categories considered in the present study are as follows:

1. Exchange a party, including its candidates.
2. Exchange a crossed-out candidate.
3. Exchange a directly selected candidate.
4. Exchange votes between directly selected candidates.

Efficiency. Figure 7 presents the average time spent by participants to cast a vote, according to the sub-tasks defined in Section 4.1.4.

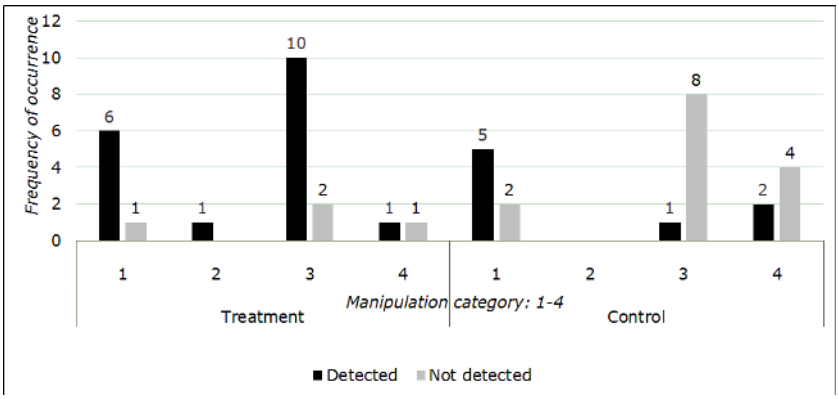

Fig. 6: Frequency of occurrence for each manipulation category.

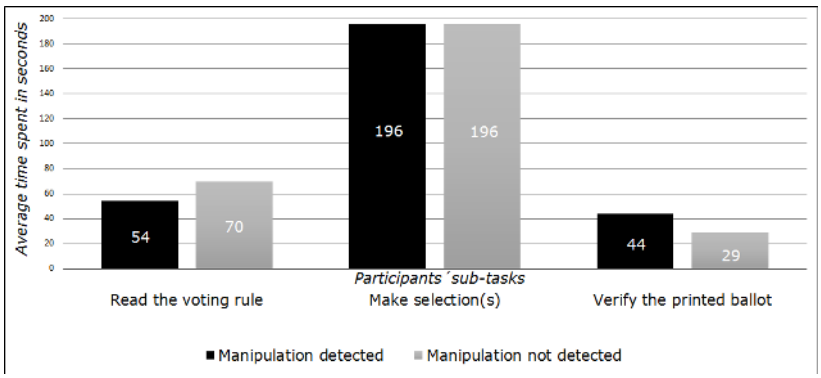

Fig. 7: Average time (in seconds) spent by participants to cast their vote.

Satisfaction. The scoring of the completed SUS questionnaires resulted in an average value of 79.94 (σ = 13.74), which, according to Sauro's normalisation method [37], can be interpreted as a grade of B. Consequently, the prototype has a high perceived usability. Furthermore, in order to analyse the impact of detecting the manipulation on the participants' satisfaction we evaluated the questionnaires separately. While the scoring of the completed SUS questionnaire from the participants that detected the manipulation resulted in a score of 80 (σ = 14.78), the scoring from those that did not detect the manipulation resulted in 79.86 (σ = 13.07). Finally, the SUS questionnaires for those in the treatment group resulted in 79.29 (σ = 13.80), and those in the control group resulted in a score of 80 (σ = 13.11).

4.5.2 Participants' perception towards verification

In this sub-section we report our findings based on a qualitative analysis of the corresponding open questions. Since our analysis was qualitative and ex-

ploratory we do not provide the number of people who mentioned a specific argument.

We evaluated participants' self-reported answers and their actual behaviour with respect to verifying the printed ballot and detecting the manipulation. In total 19 participants did not detect the manipulation. However in the post-questionnaire, 17 out of the 19 claimed to have verified their printed ballots.

Figure 8 presents participants' opinion regarding the statement *I perceived verification of the printed ballot to be a demanding task*, while Figure 9 depicts responses to: *I perceived verification of the printed ballot to be error prone.*

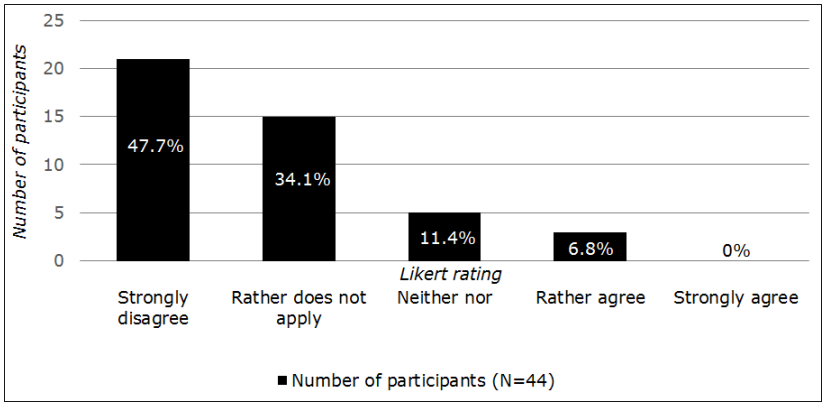

Fig. 8: I think that verifying the printed ballot is demanding.

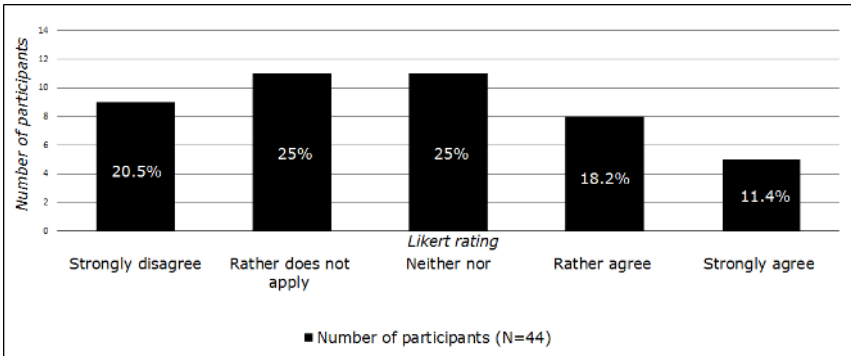

Fig. 9: I think that verifying the printed ballot is error prone.

Furthermore, by evaluating participants' responses to the question *"What motivated you to (or not to) verify the printed ballot?"* we identified five codes reflecting participants' decision to verify the printed ballot:

1. One's own attitude to verification.
2. Technology is error prone, i.e. not trustworthy.
3. Verification instructions (the reverse of the printed ballot and/or poster of the vote casting process).
4. Elections and vote integrity are important.
5. Unknown voting system.

Moreover we identified only a single code that influenced participants not to verify the printed ballot, namely *Trust in the system*. Furthermore, only one of the participants in each code category, except the first one, did not detect the manipulation. While the opposite is true for *Trust in the system*, where only one of the participants detected the manipulation. Some of the participants in the code category *One's own attitude to verify* did detect the manipulation.

4.6 Discussion and limitations

Our results show that the stimulus developed and proposed by Budurushi *et al.* [11] had a significant impact with respect to motivating voters to verify their printed ballot and thereby to detect manipulations. We could confirm the results achieved by Budurushi *et al.* [11] regarding the number of participants that verified the printed ballot and detected the manipulation, and also regarding the difference between participants' self-reported answers and their actual behaviour with respect to verifying the printed ballot and detection of manipulations.

Furthermore, the act of verifying the printed ballot and detection of the manipulation played a role in effectiveness. We identified a significant improvement in effectiveness due to the presence of the stimulus. Significantly more manipulations were detected when the stimulus was present, but no significant difference was found with respect to efficiency and satisfaction between the treatment and the control groups and between the participants that did detect the manipulation as opposed to those who did not. These findings reveal that the stimulus has a positive impact on the usability and security (verification) of the EasyVote voting scheme. In addition, 80% of the participants considering verification of the printed ballot to be easy and approximately 30% of them perceived it to be error prone. Thus, it is important to identify, analyse and minimise the factors, which, according to the participants, make the act of verifying error prone.

While analysing the different reasons that participants cited for verifying, or not verifying, their printed ballot, we found out that participants' who trust the system do not verify their printed ballot nor do they detect manipulations. Hence, it is of crucial importance for the country's democratic process to find ways to improve this before any such system is deployed in practice. One way to address this issue is to identify and deploy appropriate and effective risk communication strategies. Moreover, such strategies should increase voter awareness regarding possible risks without impairing their willingness to adopt the electronic voting system.

The limitation of the present study is that participants in this sample might be more interested in politics, being self-selecting, than the general electorate. As a result, these individuals may be more likely to verify their printed ballot and detect the manipulations than a typical voter. As such, the prevalence of verification of the printed ballot and detection of the manipulations may be biased relative to the general electorate. Moreover, we made the manipulation reporting very simple, while in genuine elections this may require strict organisational protocols, perhaps a signed form. Finally, in legally binding elections the results regarding the verification and detecting manipulations may be slightly different than in a non-legally binding election forecast. Hence, the number of voters that verify and detect manipulations might be different, depending whether the participation in the election is compulsory or free.

5 Study 2: Evaluation of the tallying process

This section describes the study in which we evaluated the tallying process. The goal of the study was to determine whether participants acting as election officials would detect discrepancies while performing the tallying process. They would be comparing the human-readable part of the ballot and the data displayed on the monitor. The task of verifying that the human-readable part of the ballot matches the data displayed on the monitor, i.e. the data stored in the QR-Code, is particularly important with respect to the integrity of the election result and therefore should be evaluated. We intentionally manipulated the QR-Codes of some ballots and randomly inserted them into the total set of ballots to be tallied. We also evaluated the usability of the prototype software according to the ISO 9421-11 standard [20].

5.1 Methodology

In this section we introduce the methodology used to achieve the goals of this study.

5.1.1 Manipulations: Introducing discrepancies

The different manipulations that were used in our study are as follows:

1. Remove votes from a candidate and assign them to another candidate.
2. Remove votes from a candidate and do not re-assign them.
3. Remove a candidate and insert another candidate instead.
4. Remove a candidate.
5. Remove a party, including its candidates.
6. Assign automatically distributed votes incorrectly.

This set includes all manipulations used in similar previous studies [14, 16].

We manipulated only six out of the 89 ballots to be tallied in order to cover all type of manipulations while minimising the probability of guessing the actual goal of the study. We randomly selected six ballots and introduced the manipulations according to a random permutation. Finally, we confronted each group with a different random order of the same six manipulations.

5.1.2 Evaluation

To evaluate the usability of the implemented tallying prototype we defined the usability components introduced in section 2.4, as follows:

Effectiveness: *The ability of participants to tally the correct election result, i.e. participants detect the introduced manipulations.* Note that we did not define any further sub-tasks to measure effectiveness, because, in contrast to the vote-casting study (Section 4), participants are trained to use the tallying prototype. Hence, by measuring effectiveness we implicitly realised our primary goal, namely to determine whether participants were able to detect the discrepancies we introduced.

Efficiency: *The time spent by participants to tally the election result.* We further measured time as follows:

1. Time spent to tally the first five ballots.
2. Time spent to manually tally the first five ballots and verify that the final tally matches the electronic outcome.
3. Time spent tallying the rest of the ballots.

We used a stopwatch to measure the time spent during the above steps.

Satisfaction: *The level of satisfaction participants experienced in tallying the election result.* In order to measure satisfaction we used a German translation of the original SUS questionnaire [6], as improved by [28].

5.2 Experimental setup, design and procedure

The experiments, including the training sessions, were carried out in our lab, which provides all the necessary equipment, e.g. a table, a laptop, a printer, chairs, monitors, and a projector. 48 participants were randomly assigned to 24 different groups of two participants. Each group had to perform the following steps:

1. Read and sign the declaration of consent for participating in the study.
2. Participate in the training workshop.
3. Tally the ballots with the implemented prototype.
4. Debrief: Reveal the actual research goal.

Participants of a group were randomly assigned to perform the following tasks, independently: (1) Operate the prototype; (2) Scan the QR-Codes. However, both participants had to verify that the data shown on the monitor matched the human-readable part of the ballot.

Furthermore, participants randomly verified the correctness of the algorithm to automatically assign votes and that of intermediate results. Hence, for each group we randomly selected 10 out of the 89 ballots, and required participants to verify all votes (directly and automatically assigned), i.e. not only those highlighted in orange. Furthermore, participants were required to randomly select five candidates on each of these ballots to verify that votes were correctly recorded. Participants compared the number of votes before and after in order to confirm that the human-readable part matched the data displayed on the monitor.

5.3 Recruiting and sampling

The participants were recruited via e-mail, advertising in social networks and flyers. 48 subjects participated (19 female, 29 male), who ranged in age from 18-54, with an average age of 25.67 years ($\sigma = 7.5$) and a median age of 24. Five participants had completed an apprenticeship; 29 participants had a high school degree or less; and 13 participants had a bachelor's degree or equivalent. One participant did not specify his/her level of education.

All participants were naïve with respect to the purpose of the study. Three different incentives encouraged participation. First, the employees of our university were interested in science and wanted to support our research. Some were psychology students, who are required by their school to participate in 30 hours of research studies. We compensated them with credit for the appropriate amount of hours. The rest of the participants were given €10 each.

5.4 Materials

In this section we list the materials used in the present study:

- The implemented tallying prototype.
- A training workshop presentation.
- 89 original electronic filled ballots from the local elections in Hesse 2011. The ballots contained directly selected candidates and/or a party.
- Five training ballots to be used during in the training workshop: Three ballots with candidates and a party selected, and two ballots that also contained crossed out candidates. Two of the five ballots required corresponding corrections by the participants.

5.5 Results

None of the first groups detected all introduced discrepancies. We thus terminated the experiment, and analysed and improved the study design.

After a systematic analysis and comparison of our study design and the tallying process in the local elections in Hesse, see 2.1.2, we identified *reading*

voters' selection(s) *out loud* during the tallying process as the only difference. Therefore, we changed our study design slightly and required participants to read voters' selection(s) out loud during the verification process. They check whether the data displayed on the monitor matches the human-readable part of the ballot during verification. This intervention significantly improved performance in comparison to our first run.

We include results from both runs of the study here. The first run consisted of four groups, which fall under the category *Silent*, as participants *silently* verified that the data shown on the monitor matched the human-readable part of the ballot. The second run consisted of 20 groups, which fell under the category *Reading out loud.* Note that due to the different group sizes, the separate evaluation regarding effectiveness, efficiency and satisfaction aims only to provide a comprehensive record of the achieved results.

Effectiveness. Figure 10 presents the ability of participants to tally the correct election result, i.e. whether the participants detected the manipulations.

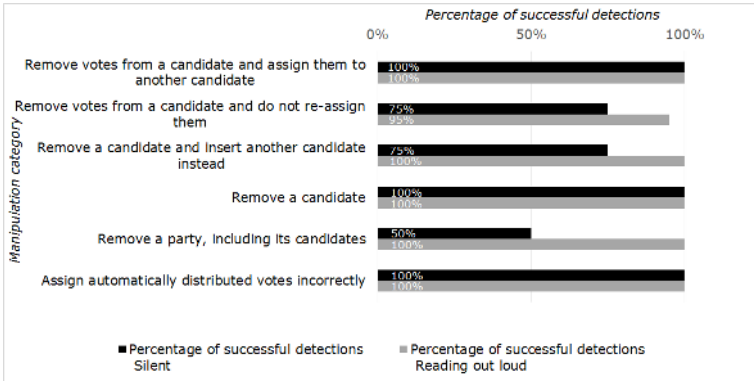

Fig. 10: The ability of participants to tally the correct election result.

Efficiency. Figure 11 presents the average time spent by participants tallying the election result according to the sub-tasks defined in section 5.1.2.

Satisfaction. The scoring of the completed SUS questionnaires resulted in an average value of 82.7 ($\sigma = 12.01$). According to Sauro's normalisation method [37] this value can be interpreted as a grade of A. Thus, the prototype is likely to be recommended. Furthermore, in order to analyse the impact of *Reading out loud* on the participants' satisfaction we evaluated the questionnaires separately. While the scoring of the completed SUS questionnaire from the silent participants resulted in a score of 86.25 ($\sigma = 5.99$). The scoring from those that read out loud while verifying resulted in 82.29 ($\sigma = 12.03$).

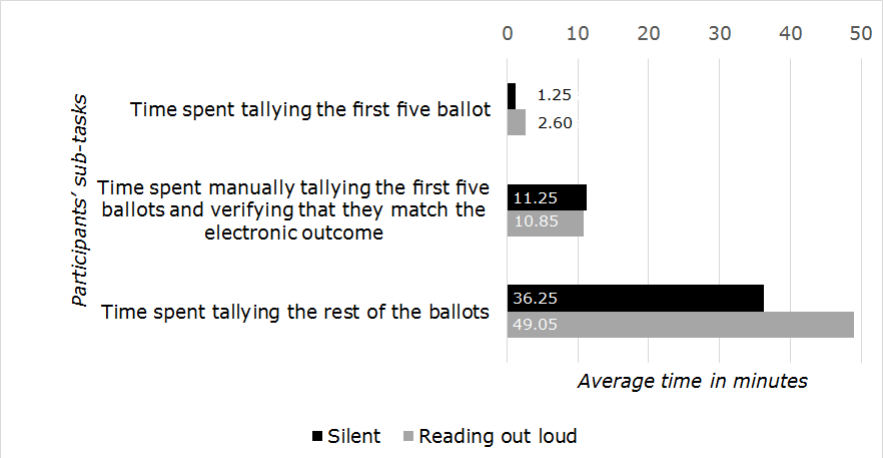

Fig. 11: Time (in minutes) spent by participants to tally the election result.

5.6 Discussion and limitations

We observed a significant increase with respect to effectiveness when the participants (electoral officials) read direct selections out loud while verifying. There were significant improvements in detecting the manipulations due to *Reading out loud*. However, due to the different group sizes, by extrapolating our results we can only assume that *Reading out loud* has no significant influence regarding efficiency and satisfaction.

There are two main limitations in the present study: First, the results regarding effectiveness might be similar, even when direct selections are not highlighted in orange, as introduced in [45]. Nevertheless, a significant difference, with respect to efficiency and satisfaction, can be assumed. Hence, in order to achieve better results with respect to effectiveness, efficiency and satisfaction, we suggest deploying both interventions (reading direct selections out loud and highlighting them in orange). The second limitation is that most of the participants in this sample were university students, and they are unlikely to be representative of the larger population (electoral officials) with respect to age and educational level. However, in legally binding elections the results regarding the verification and detecting manipulations may be slightly different, depending on whether electoral officials support similar or different electoral goals.

6 Conclusion and Future Work

In this paper we report on an evaluation of the usability of a third-generation electronic voting machine in the context of complex elections. The evaluation was carried out to asses voter and electoral official perspectives. To the best of

our knowledge, this is the first study of its kind with respect to the challenges and context it addresses.

The results of the first study, in which we evaluated the vote-casting process, show that the stimulus developed and proposed by Budurushi *et al.* [11] delivered a significant impact with respect to motivating voters to verify their printed ballot and detecting manipulations. In addition, the act of verifying the printed ballot and detecting the manipulation played a role in effectiveness. However, we identified a significant improvement with respect to effectiveness due to the presence of the stimulus. While there were significant improvements in detecting manipulation due to the stimulus presence, there were no significant difference regarding efficiency and satisfaction between the treatment and the control groups and between the participants that detected the manipulation and those who did not. These findings reveal that the stimulus has a positive impact on the usability and security (verification) of the EasyVote voting scheme. Furthermore, we observed out that participants' who trusted the system did not verify their printed ballots nor did they detect manipulations. Hence, for future work we plan to identify an adequate risk communication strategy, in order to address this sector of the population. As future work we also plan to replicate this study with the an alternative ballot design, and evaluate its impact on detecting manipulations.

In the second study we observed a significant improvement with respect to effectiveness when the participants (electoral official) read direct selections out loud while verifying. There were significant improvements in detecting the manipulations due to *Reading out loud*. However, due to the unbalanced groups, we cannot prove that *Reading out loud* has a significant impact on efficiency and satisfaction.

In conclusion, based on Neff's reasoning [31], the results of both studies show that using a stimulus *and* reading direct selections out loud makes detection of election fraud significantly more likely.

References

1. Bundeswahlgerteverordnung of 3 september 1975 (bgbl. i p. 2459), last changed on 20 april 1999 (bgbl. i p. 749). `http://www.gesetze-im-internet.de/bundesrecht/bwahlgv/gesamt.pdf` (German only), last accessed 16 November 2015.
2. A Study of Vote Verification Technology Conducted for the Maryland State Board of Elections Part II: Usability Study (2006). `http://www.capc.umd.edu/rpts/MarylandReport%202-13-06.pdf`, last accessed 16 November 2015.
3. Benaloh, J.: Simple verifiable elections. In: Proceedings of the USENIX Workshop on Electronic Voting Technology, pp. 5–14. USENIX (2006)
4. Benaloh, J.: Ballot casting assurance via voter-initiated poll station auditing. In: Proceedings of the USENIX Workshop on Electronic Voting Technology, pp. 14–21. USENIX (2007)
5. de Brettes, M.: Appareil pour voter, indiquer, autographier et contrôler les votes. Bulletin Hebdomadaire d'Association Scientifique de France **384**, 376–378 (1875)
6. Brooke, J.: SUS-A quick and dirty usability scale. Usability evaluation in industry **189**(194), 4–7 (1996)
7. Bruck, S., Jefferson, D., Rivest, R.L.: A modular voting architecture ("frog voting"). In: Towards Trustworthy Elections, pp. 97–106. Springer (2010)

8. Budurushi, J., Jöris, R., Volkamer, M.: Implementing and evaluating a software-independent voting system for polling station elections. Journal of Information Security and Applications **19**(2), 105–114 (2014)
9. Budurushi, J., Renaud, K., Volkamer, M., Woide, M.: Implementation and evaluation of the EasyVote tallying component and ballot. In: R. Krimmer, M. Volkamer (eds.) Proceedings of the 6th International Conference on Electronic Voting (EVOTE): Verifying the Vote, pp. 1–8. IEEE (2014)
10. Budurushi, J., Volkamer, M.: Feasibility analysis of various electronic voting systems for complex elections. In: P. Parycek, N. Edelmann (eds.) International Conference for E-Democracy and Open Government 2014, pp. 141–152. Edition Donau-Universität Krems (2014)
11. Budurushi, J., Woide, M., Volkamer, M.: Introducing Precautionary Behavior by Temporal Diversion of Voter Attention from Casting to Verifying their Vote. In: Workshop on Usable Security (USEC). Internet Society (2014)
12. Canard, S., Sibert, H.: Votinbox–a voting system based on smart cards. In: Workshop on e-Voting and e-Government in the UK (2006)
13. Chaum, D.: Secret-ballot receipts: True voter-verifiable elections. IEEE Security & Privacy **2**(1), 38–47 (2004)
14. Cohen, S.B.: Auditing Technology for Electronic Voting Machines. Master's thesis, California Institute of Technology and Massachusetts Institute of Technology (2005). `http://vote.caltech.edu/sites/default/files/vtp_wp46.pdf`, last accessed 16 November 2015.
15. ELECTION ASSISTANCE COMMISSION: Voluntary Voting System Guidelines 1.1 Volume 1 (2015). `http://www.eac.gov/assets/1/Documents/VVSG.1.1.VOL.1.FINAL.pdf`, last accessed 16 November 2015.
16. Everett, S.P.: The usability of electronic voting machines and how votes can be changed without detection. Ph.D. thesis, Psychology (2007)
17. Federal Constitutional Court of Germany: Decisions: Order of 03 March 2009 - 2 BvC 3/07 (2009). `http://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2009/03/cs20090303_2bvc000307en.html`, last accessed 16 November 2015.
18. Franklin, J., Myers, J.C.: Interpreting Babel: Classifying Electronic Voting Systems. In: M. J.Kripp, M. Volkamer, R. Grimm (eds.) Proceedings of the 5th International Conference on Electronic Voting (EVOTE), pp. 244–256. LNI GI Series, Bonn (2012)
19. Herrnson, P.S., Niemi, R.G., Hanmer, M.J., Francia, P.L., Bederson, B.B., Conrad, F., Traugott, M.: The promise and pitfalls of electronic voting: results from a usability field test (2005). `http://www.capc.umd.edu/rpts/Promise_and_Pitfalls_of_Electronic_Voting.pdf`, last accessed 16 November 2015.
20. International Organization For Standardization: ISO 9241-11: Ergonomics of Human System Interaction – Part 11: Guidance on Usability (1998). `https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en`, last accessed 16 November 2015.
21. International Organization For Standardization: ISO 9241-210: Ergonomics of Human System Interaction – Part 210: Human-centred design processes for interactive systems (2010). `https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en`, last accessed 16 November 2015.
22. Krimmer, R.: The evolution of e-voting: Why voting technology is used and how it affects democracy. Ph.D. thesis, Public Administration (2012)
23. Krimmer, R., Volkamer, M.: Observing threats to voters anonymity: Election observation of electronic voting. In: S.J. Krishna, N.K. Agarwal (eds.) E-Voting - Perspectives and Experiences, The Icfai University Press (2008)
24. Kuo, C., Perrig, A., Walker, J.: Designing an evaluation method for security user interfaces: lessons from studying secure wireless network configuration. Journal of Interactions **13**(3), 28–31 (2006)
25. Lindeman, M., Stark, P.B.: A Gentle Introduction to Risk-Limiting Audits. Security Privacy, IEEE **10**(5), 42–49 (2012)
26. Lindeman, M., Stark, P.B., Yates, V.S.: BRAVO: Ballot-polling Risk-limiting Audits to Verify Outcomes. In: Electronic Voting Technology Workshop/Workshop on Trustworthy Elections. USENIX (2012)

27. Loeber, L.: E-voting in the Netherlands; Past, Current, Future? In: R. Krimmer, M. Volkamer (eds.) Proceedings of the 6th International Conference on Electronic Voting (EVOTE), pp. 43–46. TUT Press, Tallinn (2014)
28. Lohmann, K., Schäffer, J.: System Usability Scale (SUS) – An Improved German Translation of the Questionnaire (2013). http://minds.coremedia.com/2013/09/18/sus-scale-an-improved-german-translation-questionnaire/, last accessed 16 November 2015.
29. MacNamara, D., Scully, T., Gibson, P.J., Carmody, F., Oakley, K., Quane, E.: DualVote: Addressing Usability and Verifiability Issues in Electronic Voting Systems. In: P. Parycek, M.J. Kripp, N. Edelmann (eds.) International Conference for E-Democracy and Open Government 2011, pp. 313–322. Edition Donau-Universität Krems (2011)
30. Mercuri, R.: Electronic vote tabulation checks & balances. Ph.D. thesis, University of Pennsylvania (2001)
31. Neff, A.C.: Election Confidence: A Comparison of Methodologies and Their Relative Effectiveness at Achieving It (2003). http://www.verifiedvoting.org/wp-content/uploads/downloads/20031217.neff.electionconfidence.pdf, last accessed 16 November 2015.
32. Pomares, J., Levin, I., Alvarez, M.R., Mirau, G.L., Ovejero, T.: From piloting to rollout: voting experience and trust in the first full e-election in argentina. In: R. Krimmer, M. Volkamer (eds.) Proceedings of the 6th International Conference on Electronic Voting (EVOTE): Verifying the Vote, pp. 1–10. IEEE (2014)
33. Pozo, J.: Implementation Project Electronic Voting Azuay Ecuador 2014. In: R. Krimmer, M. Volkamer (eds.) Proceedings of the 6th International Conference on Electronic Voting (EVOTE), pp. 47–60. TUT Press, Tallinn (2014)
34. Rivest, R.L., Wack, J.P.: On the notion of "software independence" in voting systems (2006). http://people.csail.mit.edu/rivest/RivestWack-OnTheNotionOfSoftwareIndependenceInVotingSystems.pdf, last accessed 16 November 2015.
35. Sandler, D., Derr, K., Wallach, D.S.: VoteBox: A Tamper-evident, Verifiable Electronic Voting System. In: Proceedings of the 17th Conference on Security Symposium, pp. 349–364. USENIX (2008)
36. Sandler, D., Wallach, D.S.: Casting Votes in the Auditorium. In: Proceedings of the USENIX Workshop on Electronic Voting Technology, pp. 4–4. USENIX (2007)
37. Sauro, J.: Measuring Usability With The System Usability Scale (SUS) (2011). http://www.measuringu.com/sus.php, last accessed 16 November 2015.
38. Selker, T., Pandolfo, A.: A methodology for testing voting systems. Journal of Usability Studies **2**(1), 7–21 (2006)
39. Sotirakopoulos, A., Hawkey, K., Beznosov, K.: 'I did it because I trusted you': Challenges with the study environment biasing participant behaviours. In: SOUPS Usable Security Experiment Reports (USER) Workshop (2010)
40. Stark, P.B.: Efficient post-election audits of multiple contests: 2009 california tests. In: CELS 2009 4th Annual Conference on Empirical Legal Studies Paper (2009)
41. Stark, P.B., Teague, V.: Veriable european elections: Risk-limiting audits for d'hondt and its relatives. USENIX Journal of Election Technology and Systems (JETS) **1**(3), 18–39 (2014)
42. Stark, P.B., Wagner, D.: Evidence-Based Elections. IEEE Security and Privacy **10**(5), 33–41 (2012)
43. Takaji, D.P.: The Paperless Chase: Electronic Voting and Democratic Values. Fordham L. Rev. **73**, 1711 (2004)
44. Vegas, C.: The new belgian e-voting system. In: A. Brömme, C. Busch (eds.) Proceedings of the 5th International Conference on Electronic Voting (EVOTE), pp. 199–211. Bonn: Gesellschaft für Informatik (GI) (2012)
45. Volkamer, M., Budurushi, J., Demirel, D.: Vote casting device with VV-SV-PAT for elections with complicated ballot papers. In: International Workshop on Requirements Engineering for Electronic Voting Systems (REVOTE), pp. 1–8. IEEE (2011)
46. Volkamer, M., Spycher, O., Dubuis, E.: Measures to establish trust in internet voting. In: ICEGOV 2011, Proceedings of the 5th International Conference on Theory and Practice of Electronic Governance, ICEGOV '11, pp. 1–10. ACM, New York, NY, USA (2011)

47. Vot.ar.: Vot.ar. http://www.vot-ar.com.ar/en/system-votation/, last accessed 16 November 2015.
48. Weldemariam, K., Villafiorita, A.: Modeling and Analysis of Procedural Security in (e)Voting: The Trentino's Approach and Experiences. In: Proceedings of the USENIX Workshop on Electronic Voting Technology, pp. 1–10. USENIX (2008)
49. Yee, K.P., Wagner, D., Hearst, M., Bellovin, S.M.: Prerendered User Interfaces for Higher-assurance Electronic Voting. In: Proceedings of the USENIX Workshop on Electronic Voting Technology, pp. 6–6. USENIX (2006)