Usability of Voting Systems: Baseline Data for Paper, Punch Cards, and Lever Machines

Michael D. Byrne, Kristen K. Greene, Sarah P. Everett
Department of Psychology, Rice University
6100 Main St., MS-25
Houston, TX 77005 USA
{byrne, kgreene, petersos}@rice.edu

ABSTRACT

In the United States, computer-based voting machines are rapidly replacing other older technologies. While there is potential for this to be a usability improvement, particularly in terms of accessibility, the only way it is possible to know if usability has improved is to have baseline data on the usability of traditional technologies. We report an experiment assessing the usability of punch cards, lever machines, and two forms of paper ballot. There were no differences in ballot completion time between the four methods, but there were substantial effects on error rate, with the paper ballots superior to the other methods as well as an interaction with age of voters. Subjective usability was assessed with the System Usability Scale and showed a slight advantage for bubble-style paper ballots. Overall, paper ballots were found to be particularly usable, which raises important technological and policy issues.

Author Keywords

Voting, effectiveness, efficiency, subjective usability, DRE.

ACM Classification Keywords

H5.2 Information interfaces and presentation (e.g., HCI); User Interfaces

INTRODUCTION

The debacle of the 2000 U.S. presidential election created more than just a media frenzy, it created deep concerns about how elections are conducted in this country. The statistical evidence that the infamous "butterfly ballot" produced widespread voter error is substantial [19], and images of frustrated poll workers attempting to determine voter intent from ambiguous punch cards were widely broadcast by U.S. media outlets. As a result of these and other events, in 2002 Congress passed the Help America Vote Act (HAVA), which authorized funds for the updating of voting equipment. Many jurisdictions have since adopted computer-based direct recording electronic (DRE) systems in place of older technologies such as paper ballots, punch cards, and lever machines.

While the security issues raised by this change have received considerable public attention [13], the usability issues have largely failed to capture the attention of lawmakers, the public, or the media. It appears to be implicitly assumed by vendors and policy makers that such systems must be more usable than the more traditional technologies which they are replacing, though virtually no evidence has been presented which supports such an assumption.

Fortunately, there have been several investigations into the usability of the most widely-available DRE systems [2, 5, 9]. These are important studies; however, what they do not tell us is how such systems compare to the more traditional voting methods. Particularly given the substantial concerns about the security, reliability, and considerable expense of DRE systems, it is perhaps surprising that such comparisons have not been performed. There is certainly plenty of evidence in other domains that attempts to replace paper-based technology with computer technology are not always successful [17]. Even if DREs were engineered in such a way that security, reliability, and expense were not concerns, they will still ultimately fail if there are substantial usability problems. At the very least, we should know how new systems compare to those they replace.

One of the primary barriers to performing such comparisons is that basic data on human performance with traditional technologies has never been systematically collected and disseminated. Considering the age of these technologies and their widespread use, we found the gap in

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.

the literature surprising. The research reported here is the third study in a series [6, 8] designed to help fill this gap.

Voting is an interesting domain for human factors. While the basic task of choosing *k* alternatives from *n* options (where *k* is usually 1 in U.S. elections) is not intrinsically difficult (though often comprehending the language of referenda/propositions can be vexing), expectations of privacy mean that this process will fundamentally be a human-technology interaction rather than a human-human one. The scale of the endeavor is part of what makes it challenging: literally millions of votes are cast in elections, with most of them cast all on the same day. This puts a premium on making sure votes are cast quickly, because long lines at the polls almost certainly discourage voters. Similarly, accuracy is critically important, as the consequences of elections can be substantial and long-lasting. Furthermore, voting is an infrequent task and one which always includes an influx of first-time users. This makes training an unlikely solution to any usability problems.

Of course, one of the most challenging aspects of addressing usability issues in voting is the enormous diversity of the population. People in the U.S. are enfranchised regardless of age (once past 18), income, literacy, ethnicity, language, familiarity with technology, disability, and so on. This is unusual in HCI domains, where at least literacy is almost always assumed. Issues of accessibility for a wide variety of special populations are thus an important part of the conversation, and multi-modal computer platforms capable of supporting audio input and output, pictorial displays, specialized input methods, and the like hold much promise in that regard. It is important that usability be considered for all users, regardless of their capabilities.

Another consideration which applies to voting systems which does not apply to many other user interfaces is that elections are regulated by governments in ways which many other tasks are not. In the U.S., states or counties are responsible for oversight of elections, but the federal government also has input. Currently, this takes the form of the Voluntary Voting System Guidelines (VVSG) produced by the Election Assistance Commission (EAC) in consultation with the National Institute of Standards and Technology (NIST). These are, as the name implies, guidelines that local election officials are encouraged but not mandated to follow. (See http://www.eac.gov/vvsg_intro.htm for the current VVSG.) The VVSG has an entire section devoted to usability, but notes on Volume I, page 47:

> It is the intention of the EAC that in future revisions to the *Guidelines*, usability will be addressed by high-level performance-based requirements. That is, the requirements will directly address metrics for effectiveness (e.g., correct capture of voter selections), efficiency (e.g., time taken to vote), and satisfaction. Until the supporting research is completed, however, the contents of this subsection are limited to a basic set of widely accepted design requirements and lower-level performance requirements.

This is explicit recognition of the current gap in our knowledge about voting system performance. So, if computers are going to improve the usability of the voting process, we need an answer to the question of "improvement relative to what?" How good would a DRE need to be to be better than alternative technologies? There is, of course, the possibility that a DRE could have a poor interface and actually be worse than the replaced technology. But to know this, we need clear quantitative information about the usability of pre-DRE technologies.

USABILITY CRITERIA FOR VOTING

The question of improvement naturally leads to the question of criteria. Upon what usability criteria should such a comparison be based? There are a number of possibilities for both objective and subjective measures. ISO standard 9241-11 [11] and the 2004 NIST report on voting system usability [14] recommend three metrics: effectiveness, efficiency, and satisfaction.

Effectiveness, an objective metric, is defined as the relationship between the goal of using the system and the accuracy and completeness with which this goal is achieved [10]. In the context of voting, accuracy means that a vote is cast for the candidate the voter intended to vote for, without error. Effectiveness is usually measured by collecting error rates, but could also be measured by completion rates and number of assists.

Efficiency is an objective usability metric that measures whether a user's goal was achieved without expending an inordinate amount of resources [10]. Did voting take an acceptable amount of time? Did voting take a reasonable amount of effort?

Satisfaction, a subjective metric, is defined as the user's subjective response to working with the system [10]. Was the voter satisfied with his/her voting experience? Was the voter confident that his/her vote was recorded? Measuring any kind of subjective response can be problematic and satisfaction is no exception. The NIST report [14]

recommends the use of an external, standardized instrument rather than ad-hoc queries of user's preference. Such instruments typically take the form of batteries of Likert scale questions. Standardized instruments mentioned by NIST include SUS [3], QUIS [4], and SUMI (http://sumi.ucc.ie/).

Because these measures will form the basis for future efforts at standardization, we have adopted the same set of measures in our research. It is important to note that these three measures are not necessarily related. For example, Frøkjaer and colleagues [7] both conducted a usability study and surveyed CHI proceedings and found frequent independence of these measures. Good objective performance does not always produce high satisfaction, nor are the two objective measures necessarily related.

RELATED AND PREVIOUS RESEARCH

Unsurprisingly, voting behavior has been a subject of study in political science for quite some time. Such studies have certainly included what usability researchers would consider effects of the user interface. For example, the fact that the candidate listed first on the ballot has an advantage over other candidates has been known for years [18]. More recently, various researchers have found clear evidence that different voting systems affect what is termed the "residual" vote [1, 15]. Residual votes are votes which are in some way spoiled, typically by overvoting, which is voting for more candidates than are allowed in the contest. However, sometimes undervoting, or failing to vote in a contest, may also be considered residual if there is a compelling reason to expect the voter to have cast a vote in that contest. For instance, a failure to vote for the office of president in a presidential contest is sometimes considered a residual vote. Unfortunately, these measures are somewhat indirect measures of effectiveness, since undervotes might be intentional and errors which cause the vote to go to the wrong candidate are often (though not always, see [19]) impossible to identify as errors. Measures of time and satisfaction in this literature are generally similarly indirect.

Despite this, some effects of technologies on some of these kinds of errors are obvious; it is impossible to overvote on a lever machine or a DRE. But whether or not it is easer to cast one's vote for the wrong candidate on a DRE than another technology is unclear; for example, while punch cards may have their own flaws, they do not suffer touch screen calibration problems. More direct measures of error are thus needed.

Herrson and colleagues have conducted such studies with various DREs. This research has included application of usability inspection methods [2], a laboratory study [5], and an ambitious field study [9]. While many of these results are still preliminary in that full information about methods and details of data analysis are not yet fully available, it is clear that current DREs are not yet as usable as they could be, with error rates for the presidential contest in their field study ranging from 2.5% to 4.2%. They also demonstrated differences in time taken to vote and user satisfaction (though the satisfaction measures did not use a standardized instrument). This is important research which offers substantial insight into the usability of current DREs. What this research does not provide, nor was it intended to provide, is information about how those DREs compare to older voting technologies.

We have conducted two previous studies [6, 8] focussed on that issue. In the first study [6], we evaluated three forms of paper ballot on all three of the usability criteria listed earlier. We manipulated factors such as the amount of information given to participants about the candidates and the realism of the candidates, comparing fictional candidates to projections of possible real candidates (for example, our presidential election pitted Rudy Giuliani against Hillary Clinton). The participants were not particularly representative of the electorate, being drawn from the Rice University undergraduate population, which is young and low on educational diversity while also high on cognitive ability and familiarity with technology. Thus, this pool in some sense represents a "best-case" scenario; average voters are likely to be less capable. While we found no evidence for differences in efficiency or effectiveness as a function of these factors, there was a clear effect of ballot type on subjective usability, with participants rating the "bubble" ballot (see below) as more usable. Since the bubble ballot resembles other optically-scanned forms (such as SAT exams) to which this population is often exposed, it was not clear if this preference is based simply on familiarity. Of interest also was the error rate, which was 1% per contest; 11% of the ballots contained at least one error. Given the highly-capable population, we found this surprisingly high.

In our second study [8], we used a slightly broader sample with participants being a mixture of Rice University undergraduates and people recruited through an Internet advertisement. In addition, we added a new voting method, the mechanical lever machine. We again found no effects of information condition or voting method on efficiency or effectiveness, but again found a high overall error rate: 0.9% per contest; 16% of the ballots contained at least one

error. The bubble ballot was again the high scorer on subjective usability, even though not all the participants were students.

The current study is an effort at extending these results in two ways. First, we used a more representative sample with wider ranges on demographic variables such as age and education. Second, we added a new voting method, the punch card, to the evaluation. The goal was to provide an even more detailed and accurate picture of the usability properties of these traditional technologies so that later evaluations of DREs have a basis for comparison.

METHODS

Participants

Participants were 81 local residents recruited through a newspaper advertisement. Participants were paid $25 for their participation. 2 participants declined to report their age; among the remaining 79, age ranged from 24 to 80 with a mean of 49.9 years and a median of 51 years. 37 were male and 44 were female. All had normal or corrected-to-normal vision and were US citizens fluent in English. On average, participants had voted in 7.79 national elections, ranging from zero to 28. Nine of the 81 participants had only voted nationally once before, and another six participants had no prior national voting experience. On average, participants had voted in 7.06 other types of elections (local, school, etc.).

Design

A four-factor design was used: 2 (information condition, between subjects) x 4 (voting methods, within subjects) x 2 (age, between subjects) x 4 (education, between subjects). The two information conditions used were termed "directed" and "undirected," to which participants were randomly assigned. In the directed condition, participants were instructed as to which candidates and propositions they were vote for (and given no other information about the candidates). In the undirected condition, participants were given a guide describing the various candidates and propositions and told to vote as they would if these were the actual candidates and propositions they faced in a real election. This guide was modeled after those produced by the League of Women Voters, and participants were free to choose how much time they devoted to it. Participants were allowed to mark the guide and have it with them when voting if they so chose. The primary motivation for this manipulation is that it is clearly easier to conduct mock election studies in which users are simply instructed who to vote for, that is, where the actual decisions about who vote for are decided by the experimenter. In real elections, however, voters can decide for themselves who to vote for and may do so without making the decision in advance. While our undirected condition may not be a perfect analog to a real election, we wanted to assess whether added realism influenced performance.

There were four voting methods used in this experiment: arrow ballot, bubble ballot, punch card, and lever machine. All participants voted with all four methods, and order of presentation was counterbalanced.

The other two factors were demographic variables of age and education. Age was divided by a median split, which occurred at age 51. Because two participants opted not to include their age, they were excluded from these analyses.

Education was based on a self-report measure. The original measure consisted of six categories: did not complete high school, high school diploma, some college, Bachelor's degree, Master's degree, and Doctoral degree. Due to small numbers of respondents in the most extreme categories, they were combined with the next most extreme to yield four categories: high school or less, some college, Bachelor's degree, and graduate degree. Table 1 shows the frequency for each category; one participant did not report their level of education.

	Frequency
High school or less	15
Some college	38
Bachelor's degree	17
Postgraduate degree	10

Table 1. Frequency of each educational level

We also collected data on other demographic and experiential factors such as computer experience and number of elections previously voted in with each technology.

We had three main dependent variables: errors, ballot completion time, and subjective usability. In the case of the directed condition, errors were defined as deviations from the list of candidates given to the voter. In the undirected condition (in which participants made their own selections), participants were instructed to vote the same on all four ballots and then interviewed about their selections after completing all four ballots. Errors were then defined as inconsistencies in response. For example, if a participant

voted for Bob on three of the four ballots and reported voting for Bob in the interview, but voted for Mary (or nobody) on one ballot, then the vote on that ballot was scored as an error. Error rate for each ballot was then computed for each participant and this used as the dependent measure.

Ballot completion time was measured with a stopwatch, which was started when the participant entered the voting booth and stopped when s/he exited.

Subjective usability was measured with the SUS [3]. The SUS contains 10 5-point Likert scales related to various aspects of usability. The ratings for the items are combined to produce a single usability score. Scores on this scale range from 0 to 100 with higher scores indicating higher perceived usability.

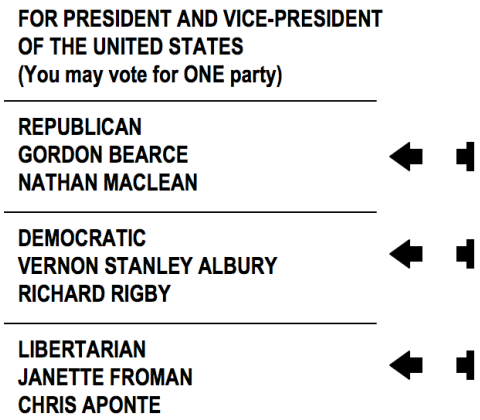

Figure 1. Arrow ballot

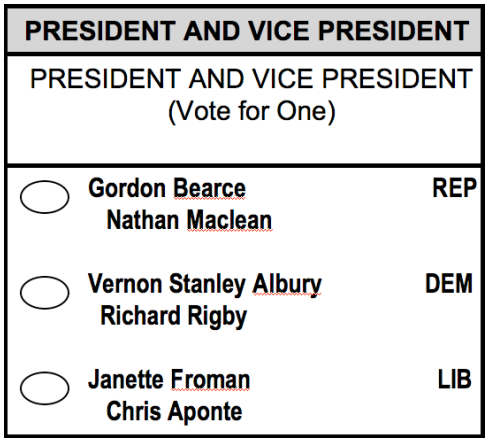

Figure 2. Bubble ballot

Materials

Both types of paper ballot were based on actual optical-scan ballots currently in use in U.S. elections. For the arrow ballot, depicted in Figure 1, participants indicated their choice by drawing a simple line to complete an arrow. For the bubble ballot, depicted in Figure 2, participants indicated their choice by filling in an oval, or bubble.

The punch card stations were VotoMatic III stations (see Figure 3) where participants turned pages to see various contests and used a metal-tipped stylus to punch a hole to indicate their choice. The punch card stations were purchased at public auction from Brazoria County, TX.

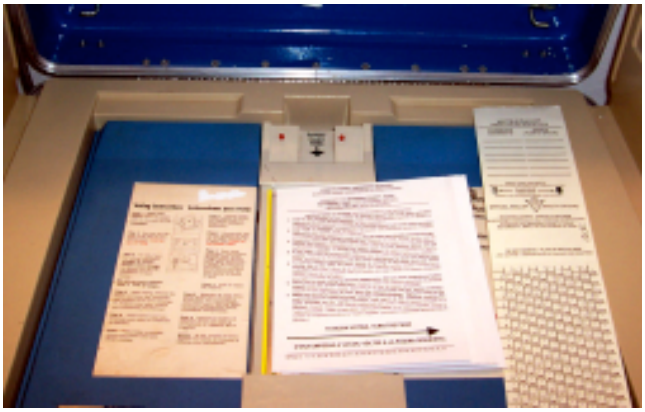

Figure 3. Punch card station

The mechanical lever machines, pictured in Figure 4, were manufactured by Automatic Voting Machines, Inc. and purchased at public auction from Victoria County, TX. Participants indicated their selection by depressing the small mechanical lever next to the label for their choice.

Figure 4. Mechanical lever machine

Each voting method used the candidates and propositions used in our previous experiments. Candidate names were produced by a publicly-available random name generator found at http://www.kleimo.com/random/name.cfm with

the "obscurity factor" set to 15. The 21 offices ranged from national contests to county contets, and the six propositions were real propositions that had previously been voted on in other counties/states and that could easily be realistic issues for future elections in Harris County, TX, the county in which this study was conducted. Because laws on straight-ticket voting (that is, casting a vote for the same party in each contest) vary widely in the U.S. (some states require it, some states forbid it) and it is difficult to represent all of the alternatives, participants were not given a straight-ticket option.

In the directed condition, there were three types of candidate lists. The mixed list contained an approximately equal mix of Republicans and Democrats. The primarily Republican list contained 85% Republican candidates, while the Democratic list 85% Democrats. All lists instructed participants to vote "yes" for four propositions and "no" for two.

Disclaimers were put on the guides, instructions, and debriefing that reminded participants that the materials and information in the guide were strictly for research purposes and may not accurately reflect the views of any real person.

Procedures

Participants first gave their informed consent and then read the instructions for the experiment. Participants were instructed to vote the same way on all four ballots. Those in the directed condition were given their list; those in the undirected condition were given their guide and allowed to examine it for as long as they wanted before voting. Participants then voted with all four voting methods (order was counterbalanced across participants). For participants in the guide condition, then were then interviewed about their choices. All participants were then given a survey which included demographic information, SUS ratings for all four voting methods, and some other questions regarding their past voting and computer experience. Ballots on the lever machine were scored immediately while the other ballots were scored later.

It is important to note that while our paper ballots were modeled after ballots designed for use in optical scanners, the ballots were not scored by scanner but were scored by hand by one of the authors. In some cases participants did not carefully follow the instructions printed on the ballot and marked them in a way which clearly indicated intent but would probably not have registered properly on a scanner. Similarly, recording of the selections on the lever machine was done by hand by one of the authors rather than relying on the lever machine counters, and scoring of the punch cards was also done by hand and not by a punch card reader. Thus, the error measures represent something of a best-case scenario and are lower than what would be observed in a mechanical vote count, though they probably correspond well to a hand count.

RESULTS

The broad sample produced data with several outliers on the objective measures. Six participants had their data removed from the error analysis due to abnormally high error rates, defined as more than 15% errors on all four voting methods. Similarly, seven participants (not all the same participants) were not included in the analysis of ballot completion times for having one or more times more than 3.5 5% trimmed standard deviations from the group mean. There were no differences according to which type of list was used in the directed condition so this was treated as a single condition. Except where noted, the analysis was a four-way (2x4x2x4) mixed-model factorial ANOVA, with factors and levels corresponding to the specification in the Design section. *p*-values were adjusted by Greenhouse-Geisser correction when appropriate.

Errors

By far the most compelling results concern error rates. There are multiple ways error rates can be computed; we considered errors by contest (or race) and errors by ballot. Per-contest error rate was a function of information condition and voting method (presented in Figure 5). Unlike in our previous experiments, there were clear differences between the voting methods, $F(3, 174) = 3.90$, $p = .027$. In addition, participants in the directed condition made reliably fewer errors than those in the undirected condition, $F(1, 58) = 4.42$, $p = .04$. This suggests that other studies based on directed voting may be underestimating the real error rate.

Per-contest error rate as a function of voting method and age are presented in Figure 6. While there was no main effect of age, there was a reliable interaction between age and voting method, $F(3, 174) = 9.00$, $p < .001$. This interaction is complex; while there is no reliable effect of age for the paper ballots, this is not the case for the other methods. Older participants had a higher error rate with the lever machine, but a lower rate with the punch cards. Because the voting method used in the area where the experiment was conducted used punch cards before converting to DREs, and lever machines were never used there, one possible explanation for this difference is experience. However, error rate with the punch card did not correlate with the number of elections people had voted with punch cards previously, nor did lever machine

experience correlate with lever machine error rate. This suggests some factor other than experience is responsible for these differences.

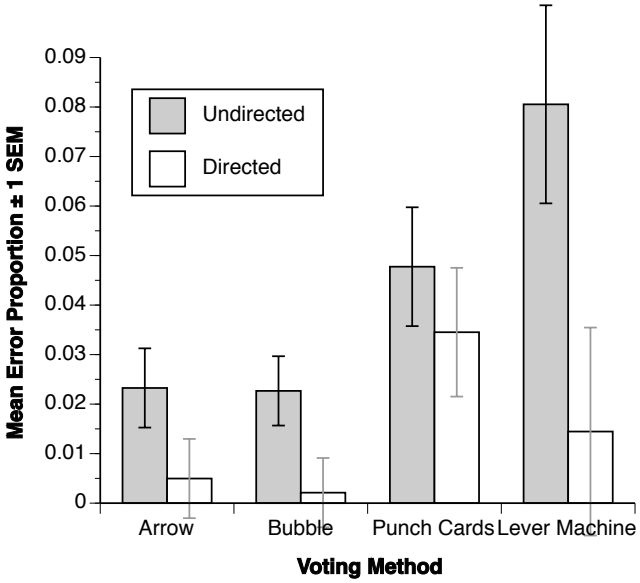

Figure 5. Per-contest error rate vs. voting method and information condition.

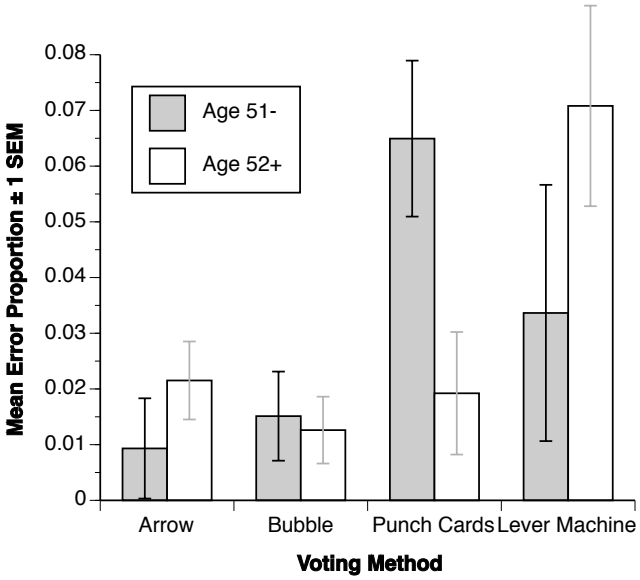

Figure 6. Per-contest error rate vs. voting method and age.

We also examined ballot position effects; this is, were error rates higher for contests further down the ballot? The first 21 ballot positions only were examined because the final six were proposition and were on a second page on the paper ballots; data are presented in Figure 7. The effect of position is not reliable, nor is the linear trend. (For this analysis, individual votes rather than ballots were taken as the unit being measured.) So, while the "importance" of the earlier contests was in some sense higher, there is no evidence that participants devoted less attention to the later contests.

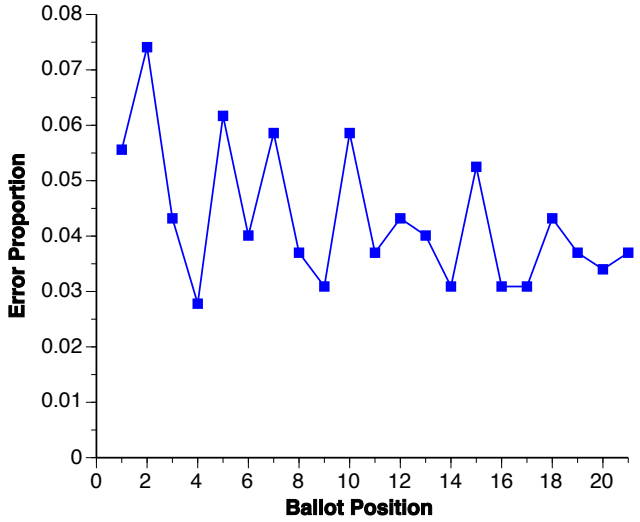

Figure 7. Per-contest error rate vs. ballot position.

Finally, another way to consider error rates is by ballot. A ballot either does or does not contain at least one error. Frequencies by voting method are presented in Table 2. A chi-square test of association revealed no reliable effect of voting method ($X^2(3) = 0.76$, $p = .86$), that is, voting method did not appear to affect the frequency with which ballots containing errors were generated. 78 of the 300 (26%) ballots collected contained at least one error.

Errors			
	None	**At least 1**	**Total**
Arrow	56	19	75
Bubble	58	17	75
Punch	54	21	75
Lever	54	21	75
Total	222	78	300

Table 2. Ballot error frequencies by voting method

The fact that the higher error rates per contest for punch cards and lever machines did not produce higher error rates per ballot suggested that errors are not independent; that is, participants who made one error were prone to make

further errors. The different voting methods did not appear to influence how likely any single participant was to err on a particular ballot, but rather influenced how many errors they made once they did.

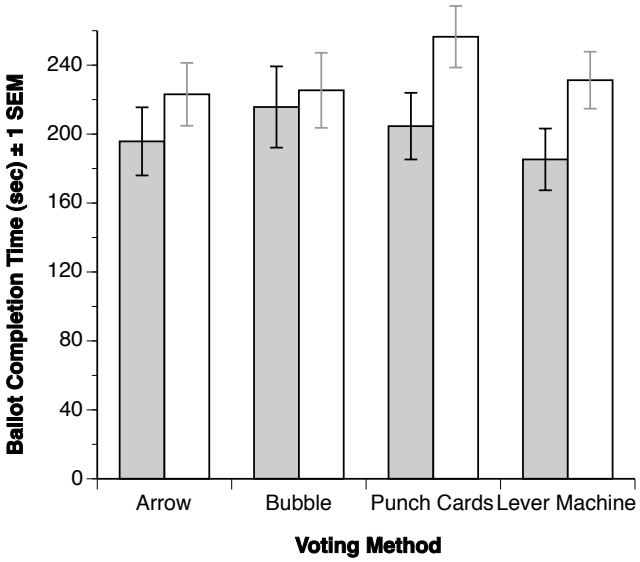

Figure 8. Ballot completion time vs. voting method and information condition. Gray = undirected, white = directed.

Ballot Completion Time

Results for ballot completion time as a function of voting method and information condition are presented in Figure 8. As in our previous experiments, voting took approximately the same amount of time regardless of voting method. However, as shown in Figure 9, there was an overall effect of education on ballot completion time, with higher levels of education associated with more rapid ballot completion, $F(3, 57) = 3.05$, $p = .036$; linear trend $p = .019$. While this is not a particularly counterintuitive result, the actual mechanism underlying this effect is unclear. Perhaps this is simply a function of familiarity, as more-educated users do more indicating of choices using technology, but the fact that this effect did not show up in the error rates nor interact with age suggests that the underlying explanation may be more subtle.

None of the effects of voting method, information condition, or age were statistically reliable, nor were there any interactions. While consistent with previous results, the lack of effect of information condition is interesting; giving participants a list of selections to make did not appear to speed them up.

Subjective Usability

Figure 10 depicts the mean SUS rating as a function of voting method and information condition. As in our previous experiments, the bubble ballot scored the highest. The main effect of voting method was reliable, $F(3, 192) = 4.85$, $p = .003$. Individual t-tests confirmed that the bubble ballot was rated significantly higher than the arrow ballot and lever machine. There were no effects of age or education nor were there any reliable interactions, though there was a hint that the undirected condition produced higher SUS scores than the directed condition, $F(1, 64) = 3.59$, $p = .063$.

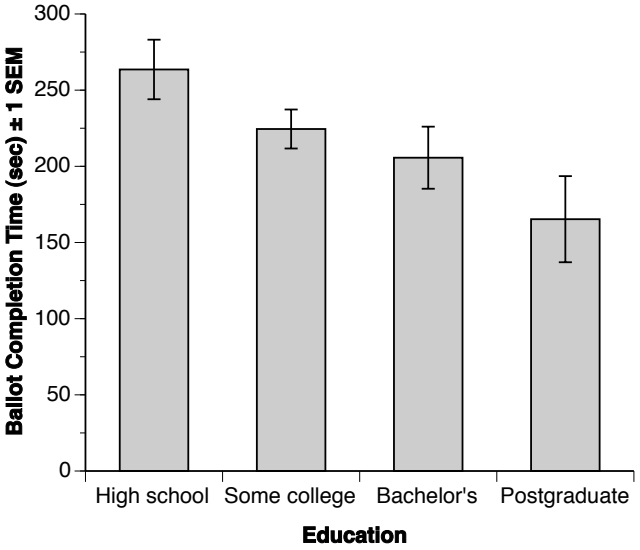

Figure 9. Ballot completion time vs. education

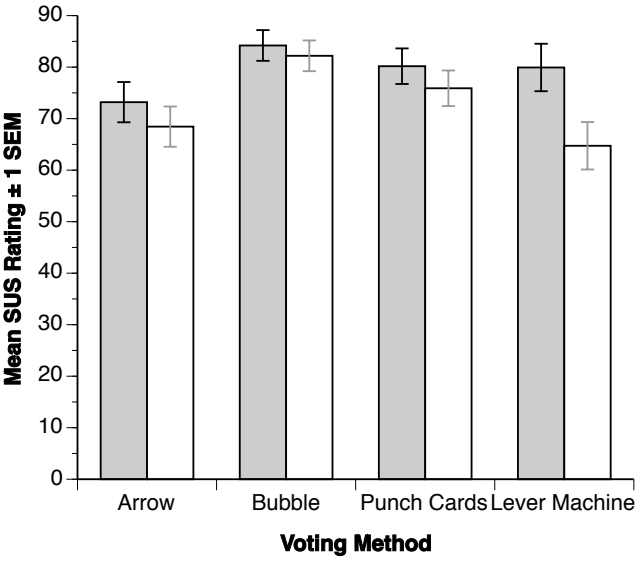

Figure 10. Subjective usability score (SUS) vs. voting method and information condition. Gray = undirected, white = directed.

Interestingly, SUS ratings did not correlate with objective measures of performance. That is, error rate and ballot completion time did not reliably predict SUS ratings.

Furthermore, participants were asked to report their level of computer experience and number of hours of computer use each week; none were meaningfully correlated with any of the dependent measures. "Computer experience" may well be correlated with performance on electronic voting machines, but it was not for our tasks.

DISCUSSION

These data serve as a reference point for future discussions concerning the usability of DREs vs. other technologies. Paper ballots appear to have generally good usability; overall error rate for paper was just over 1.5%, which is not only lower than the error rate for lever machines and punch cards, but lower than the rates found for commercial DREs [9]. This comes at no cost in terms of efficiency or subjective usability, at least relative to punch cards and lever machines. Furthermore, with respect to effectiveness, paper ballots seem less sensitive to the effects of age, which were substantial for punch cards and lever machines. The age effects suggest differential enfranchisement of different subpopulations, which is generally considered counter to the aims of election systems. The relationship between age and efficiency for DREs is still unknown.

The fact that paper fared well is not necessarily surprising. In many settings and for many tasks, paper has excellent usability [17]. Most people have a great deal of experience dealing with paper and we as a culture have thousands of years of it. With paper, there are fewer opportunities for failures due to unclear instructions, indirect mappings between actions and candidates, inappropriate configuration, and other problems which Roth [16] has described for lever machines and punch cards.

On the other hand, paper ballots do have disadvantages. For example, paper ballots are not particularly usable for those with visual impairments or a variety of physical disabilities. This is a serious drawback for paper and one of the potential major advantages for DREs. Thus, an important goal for future research is to determine the actual usability of systems designed for those with disabilities.

Furthermore, all of the technologies which have been systematically evaluated so far require that the voter be able to read; it is fair to assume that the illiterate will not fare well with such technologies, but there is not clear data on this question.

The fact that there were no differences in ballot completion time between the four voting methods suggests that the time taken to vote is dominated by processes which do not heavily depend on the details of the interaction. The effect of education suggests that the major driver of ballot completion time is simply time taken to read and make decisions, which applies regardless of voting method.

It is our hope that these findings will help inform efforts to establish usability standards for voting systems such as the one by NIST [14]. Of particular interest for this purpose were the effects of information condition. It will certainly be easier to perform summative usability testing by vendors or independent testing agencies using a directed condition; however, our data indicate that this may suppress error rates and possibly subjective usability ratings. More data on this question would certainly be welcome to help clarify this issue.

The findings on accuracy have important public policy ramifications. So far in our studies, we have not seen a voting technology with an error rate under about 1% (even for a highly advantaged population [6]). Even if tabulation and canvassing procedures are completely error-free (which seems unlikely), our ability to measure intent has limited precision. This has implications for how we think about tiebreaking and recount procedures. How sure can we be about the results of any election with a margin of victory of less than 1%?

Obviously there are still an enormous number of unanswered questions. Both our studies and the work by Herrnson and colleagues have drawn from English-speaking populations. Whether these results generalize to subpopulations who do not read English but are literate in other languages such as Spanish or Chinese is as yet unknown. However, we have translated all of our materials into both of those languages and are currently collecting data on this question.

The shift to DREs has created other usability issues as well. Largely due to security concerns, 27 states now require that DREs produce a paper record of a voter's choices which the voter can inspect before casting their ballot. How this affects usability is not yet completely clear. These "paper trails," often called VVPATs (for voter-verified paper audit trail), are typically printed in fairly small print with inexpensive thermal printers. Names are typically not familiar, easy-to-read bits of text. Thus, there is reason to be concerned about how often voters actually verify their selections. Furthermore, how accurate are they when they do so? If they do verify selections, this will clearly add additional time to the voting process—but how much? Does this affect the perceived usability of the systems? A second issue here is how easy it is to work with these paper audit trials once they have been generated. If a hand count

is required, how long does it take to perform a count from spools of thermal paper, and how accurately can people perform this count?

This raises the more general issue of usability of procedures and technologies not just for voters, but for poll workers and other election administrators as well. Poor usability may influence the accuracy and security of post-election activities such as tabulating and canvassing. This is a largely unexplored problem.

So, while this experiment and other studies make some progress toward the goal of understanding the role of usability in elections, there is still much work to be done.

ACKNOWLEDGMENTS

We would like to thank Phil Kortum and Stephen Goggin for comments on and assistance with this research. This research was supported by the National Science Foundation under grant #CNS-0524211 (the ACCURATE center). The views and conclusions expressed are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of the NSF, the U.S. government, or any other organization.

REFERENCES

1. Ansolabehere, S., & Stewart, C., III. (2005). Residual votes attributable to technology. *Journal of Politics, 67 (2)*, 365–389.
2. Bederson, B. B., Lee, B., Sherman, R. M., Herrnson, P. S., & Niemi, R. G. (2003). Electronic voting system usability issues. In *Human Factors in Computing Systems: Proceedings of CHI 2003* (pp. 145–152). New York: ACM.
3. Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London: Taylor and Francis.
4. Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of SIGCHI '88* (pp. 213-218). New York: ACM.
5. Conrad, F. G., Lewis, B., Peytcheva, E., Traugott, M., Hanmer, M. J., Herrnson, P. S., et al. (2006). *The usability of electronic voting systems: Results from a laboratory study.* Paper presented at the Midwest Political Science Association, Chicago, IL. April 2006.
6. Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: Efficiency, effectiveness, and satisfaction. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.
7. Frøkjaer, E., Hertzum, M., & Hornbaek, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Human Factors in Computing Systems: Proceedings of CHI 2000* (pp. 345–352). New York: ACM.
8. Greene, K. K., Byrne, M. D., & Everett, S. P. (2006). A comparison of usability between voting methods. *Proceedings of the 2006 USENIX/ACCURATE Electronic Voting Technology Workshop*. Vancouver, BC, Canada.
9. Herrnson, P. S., Niemi, R. G., Hanmer, M. J., Bederson, B. B., Conrad, F. G., & Traugott, M. (2006). *Voters' abilities to cast their votes as intended.* Paper presented at the Workshop on the Usability and Security of Electronic Voting System.
10. Industry Usability Reporting Project. (2001). Common industry format for usability test reports (ANSI/INCITS 354-2001).
11. International Committee for Information Technology Standards. ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDT)s part 11. Guidance on usability.
12. Kimball, D. C., & Kropf, M. (2005). Ballot design and unrecorded votes on paper-based ballots. *Public Opinion Quarterly, 69(4)*, 508–529.
13. Kohno, T., Stubblefield, A., Rubin, A. D., & Wallach, D. S. (2004). *Analysis of an electronic voting system.* Paper presented at the 2004 IEEE Symposium on Security and Privacy, Oakland, CA. May 2004.
14. Laskowski, S. J., Autry, M., Cugini, J., Killam, W., & Yen, J. (2004). Improving the usability and accessibility of voting systems and products. NIST Special Publication 500-256.
15. Mebane, W. R. (2004). The wrong man is president! Overvotes in the 2000 presidential election in Florida. *Perspectives on Politics, 2(3)*, 525–535.
16. Roth, S. K. (1998). Disenfranchised by design: Voting systems and the election process. *Information Design Journal, 9(1)*, 1–8.
17. Sellen, A. J., & Harper, R. H. R. (2001). *The myth of the paperless office*. Cambridge, MA: MIT Press.
18. Taebel, D. A. (1975). The effect of ballot position on electoral success. *American Journal of Political Science, 19(3)*, 519–526.
19. Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Herron, M. C., & Brady, H. E. (2001). The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review, 95 (4)*, 793–810.