General Ergonomics
13414 Blackstone
Universal City, TX 78148
genergo@gmail.com

August 7, 2006

To whom it may concern:

This letter addresses the usability testing conducted on Vote-PAD/DESI and Vote-PAD/Hart systems on July 19 and 20, 2006. Documents reviewed are listed in Appendix A.

Usability testing refers to the systematic evaluation of the "interaction between people and the products, equipment, environments, and services they use" (McClelland, 1990). The purpose of usability testing is to determine the effectiveness, acceptability and ease-of-use of a product with a specific user population. Usability testing is often done to identify problems, as well as design-related solutions to those problems. Although usability testing is often done in a laboratory, accurate results, that can be generalized to others, depends on testing being done under conditions as similar as possible to those in an actual user situation – this includes all of the typical stressors. This is known as "ecologic validity" – how close the test environment and testing tasks resemble the actual environment and conditions in which the product will be used. There are a number of issues in the testing conducted on July 19 and 20 that may raise concern in terms of contextual and procedural bias:

1) **Testing should resemble the actual situation as closely as possible**.

    a. That is, participants should complete voting in a simulation that closely resembles normal voting. Adding additional tasks, such as skipping sections, doing additional write-in votes, etc. gives additional information, but does not answer the essential question: Can this person use this device/process in the way it was intended to be used, with the intended consequences, easily? At this point, it is unclear whether the intended audience can vote using the products tested, as usability testing that accurately depicts a voting scenario was not conducted.

    b. It is not readily apparent why the additional tasks were added (during testing the participant was required to vote four write-ins, skip a race twice, continue voting and go back to the race). One valid reason for doing this would be if an assessment found that most California voters, in a given voting session, vote for four write-ins, skip a race twice, continue voting and go back to the race. This 'proof of validity' (offered in the testing report) for inclusion violates human factors usability precepts "it should be noted that there is nothing within this system that prevents a voter from choosing to back up and vote a previously skipped contest and, therefore, it was valid to test such a situation."

c. Additional tasks, such as skipping sections, doing a particular number of write-in votes, etc. introduces an additional set of questions and opportunities for mistakes that has nothing to do with the individual's own ability to vote and check their own answers.

d. Additional tasks prolonged the testing scenario, possibly introducing fatigue and additional frustration for participants, which could have had an impact on the number and pattern of errors. Indeed, it limited the verification information available. According to the report "unfortunately, after taking such a long time to vote their ballots, most of the voters chose to decline that verification step after voting, openly expressing their concern over the time involved".

e. Phase 2: "Verification of the Ballot". Unless a voter normally has to look at another person's ballot and determine how that other person voted, this material is irrelevant to this situation. If the purpose is to determine whether the individual can ascertain how they, themselves, voted and whether they voted too many times in a single contest, then they should have tested their own vote, to be certain it was done as they intended, during Phase 1. Unfortunately, fatigue prevented several participants from doing this. Phase 2 might have validity if the purpose is to determine whether blind or low vision individuals can be hired to check other voters' responses for them and give them feedback on their voting accuracy.

f. If an individual familiar with the new product/procedure would be available during voting, this person should also be available during testing.

2) **Worst case scenario testing reveals worst case information**. At times it is beneficial to use very difficult tasks during usability testing, sometimes even using a "worst case scenario". There are two reasons for doing this; first, by introducing difficult tasks the maximum capabilities of the participants can be defined. This type of testing is often done when it is imperative to design a task within a persons capabilities, in order to reduce human error. An example would be testing airplane pilots on dual task performance; as their primary task becomes more difficult, they spend less time on secondary tasks. By carefully annotating where and when this happens, designers gain knowledge about designing the equipment and tasks in a cockpit so the pilot is not over-taxed. The second reason for this type of testing is to identify the maximum number and diversity of problems associated with a product or procedure. This is important so that designers can use the information to re-design the product or procedure to address the identified problems. The difficulty can be in the interpretation of this information. While a carefully designed study that slowly introduces more and more difficulty can tell you of a participants basic capabilities, a study that simply has a participant do very difficult tasks does not answer the same question. For example, if a person can lift and carry 50 lbs (do the most difficult task scenario), she or he can probably lift and carry 30 lbs (accomplish the less difficult 'basic' task). However, if testing shows the participant cannot lift and carry 50 lbs (do the most difficult task scenario), the tester has no idea if they can lift and carry 30, 25 or even 20 lbs (accomplish the less difficult 'basic' task). In other words, using a worst case scenario does not answer the question of whether the

participant can do a particular job or task, other than the one tested. In this testing situation (Vote-PAD/DESI and Vote-PAD/Hart), the testing scenario appears to be more difficult than a normal voting situation due to the additional tasks of skipping sections, going back, doing four write-ins, etc. This helps identify additional errors and potential solutions[1], but it does not tell you whether the participant can vote and check the accuracy of their own votes. Therefore, the information is excellent for designers and those who will develop solutions for identified problems, but it is less useful for making decisions about the benefit of the product or device during normal voting.

3) **Usability testing should be unobtrusive**. As much as possible, usability testing should be invisible to the user. For example, camera set up should be done before arrival and tested, so that participants can act as they normally would during voting. The only adjustments should be to capture the full individual and their movements on camera. For example, if a camera must be raised or lowered if the person is in a wheel-chair, taller or shorter than the cameras' original settings. In addition, the position assumed by the participant should be the actual position that is used during normal voting. This also means that the monitors conducting the testing should offer no coaching, no additional instructions during the task (unless those instructions are part of the normal voting process), and no feedback to the participant. In additions, no additional distractions (other than those that would be present in a normal voting situation) should be present.

4) **In testing, all instructions need to be precise and exactly the same for each participant**.

    a. This is listed as a limitation of the testing (According to the report, "There were changes during testing both in the steps executed by the monitors and in the instructions given by the county employees acting in the role of poll workers"), but those conducting the testing state they do not believe it influenced the results. They do not explain why this would not influence the results, as it is a basic tenet of all research and data collection, as to do otherwise can bias the results. The report does not include a section on whether an attempt was made to control this influence by ensuring that an equal number of participants from each disability group were briefed by the same person. That is, if each person giving instructions gave them to an equal number of persons using each device and an equal number from each disability group, then this potential influence on the outcome would have been controlled.

    b. "Instructions" during research include all instructions on how to use a product and do a procedure. It also includes any verbal feedback to participants. This means that all feedback to participants should be exactly the same. Positive feedback, negative feedback and coaching during testing have been shown to influence the

[1] For example, if the poll workers explaining the procedure and device had difficulty making the intentions and use of the product clear during testing, one potential solution could be to place poll workers who are well-educated in product use at convenient voting locations. This would be one way to use the larger number and diversity of problems identified by this type of testing.

test results. Therefore, they must all either be absent or identical for each participant.

5) **Usability testing should be free of bias, even for those conducting testing**. In the instructions to those conducting the testing (#9), it states "Review the vote pad booklet to determine if you can see how the voter actually voted in any of the contests". This wording biases the individual doing the testing, as it seems to indicate that this will be a problem – that is – it is likely the individual doing the testing will not be able to determine how the voter actually voted.

6) **Measures must reflect the target audience, the product, and the actual situation in which the person would act.**

    a. Although it may be true that staffing is an issue at some voter locations, this should not mean that time-to-vote becomes a criteria for a person with a disability, as some 'accommodations' require the addition of time to complete a task. Also, what is the "appropriate" amount of time; to what will this measure be compared?

    b. It is unclear what proportion of voting errors, in a normal election, are write-in errors; but clearly, they comprised the vast majority of errors in this test. It is difficult for the reader to make a clear and logical decision regarding the results, without knowledge of how often voters use write-ins and how many errors occur during write-ins. According to the report "fifty-five of the one-hundred ballot errors that occurred were related to write-in voting".

7) **Ease-of-use is partially determined by user feedback**. Acceptance testing must be accomplished, in part, by having participants report on how easy a product was to use and what problems they had. However, if they have no reference point of comparison, their feedback cannot be taken in context. It is not clear from the report whether any of the participants had voted previously and how this experience differed from former experiences. Although the exit surveys are included, there is still no organized report on how this experience compares with previous voting experiences. This information would seem essential to determine if the present product/procedure offers advantages or disadvantages over traditional voting.

8) **Concomitant verbalization is a good technique during usability testing** (verbalizing what is being done out loud as an action is done). The purpose of this technique is to have the participant "think out loud", so the tester can understand *why* a process or device is a problem or one is better than another. Without this information, an evaluation may discover that a mistake has been made (an error), but not why. That is, they may not understand whether the product design or an errant thought process might have contributed to the error. Verbalizing what you are doing, while you are doing it, requires additional cognitive effort as this is normally a process that one does automatically. The participant must be permitted to do the task and verbalize what they are doing and why, without interruptions. Additional instructions, coaching or feedback will disrupt the

process. If the participant has to listen to and process additional feedback, while verbalizing what they are doing, they are likely to lose track of what they are doing as they seek to listen to, remember, and act on the new instructions. Therefore, as previously stated, this technique requires the monitor to quietly observe (or film) without disrupting the process or distracting the participant.

9) **Each process should clearly be evaluated, as to the impact on the system**. The report appears to assume a relative equality of instructions (staff, Braille, audio). It would be helpful to have the instructions evaluated separately for clarity, understanding, need for repeated exposure (looking back, asking questions), and number of errors during voting.

10) **The target audience needs to be well-defined and appropriately represented**. This is necessary for accuracy of representation and generalizability of results. Although the data was presented for several disability categories, talking about a certain percentage of participants that had difficulties, such as 1 of 3 with developmental disabilities, or 1 of 8 who were blind, leads the reader to false conclusions. Statistical assessments can not reach an appropriate level of significance; instead the reader is left to make conclusions from insufficient descriptive data. It is possible to conduct assessments with a low number of participants; if repeated measures testing is done (the same person is tested under different conditions). However, this was not done.

In summary, additional testing which accurately depicts a voting situation appears necessary to ascertain the potential benefits or difficulties associated with this product and procedure. Any testing should follow standard usability testing guidelines and research procedures.

I hope this information is informative regarding human factors usability testing. If I can be of further assistance, please do not hesitate to call me (1-210-391-8000).

Valerie Rice
PhD, CPE, OTR/L

Reference:

McClelland, I. (1990). Product assessment and user trials. In J.R. Wilson and E.M. Corlett (Eds)., Evaluation of human work: A practical ergonomics methodology (p. 219). New York: Taylor & Francis.

Appendix A

Reviewed:

1) The monitor procedures.
http://www.ss.ca.gov/elections/voting_systems/procedures_sos_monitor.pdf

2) Monitor Record for the first day.
http://www.ss.ca.gov/elections/voting_systems/monitor_records_day1.pdf

3) Consultant's report.
http://www.ss.ca.gov/elections/voting_systems/vote_pad_consulting_report_final.pdf
.
4) Staff Report. http://www.ss.ca.gov/elections/voting_systems/votepad_staff_report_final.pdf

5) Concerns. http://www.vote-pad.us/Media/CACertTestResponse.asp
http://www.vote-pad.us/Media/CertificationTestingObservations.pdf

6) Additional information listed at: http://www.ss.ca.gov/elections/elections_vs.htm