Overview of Proposed Human Performance Metrics for Voting Equipment

Version date: March 10, 2006

---

1. Introduction

TGDC Resolution #05-05 reads in part:

> "Title: Human Performance-Based Standards and Usability Testing
>
> The TGDC has concluded that voting systems requirements should be based, wherever possible, on human performance benchmarks for efficiency, accuracy or effectiveness, and voter confidence or satisfaction. ... Conformance tests for performance requirements should be based on human performance tests conducted with human voters as the test participants. ... Therefore, the TGDC directs NIST to:
>
> 1. Create a roadmap for developing performance-based standards, based on the preliminary work done for drafting the standards described in Resolution # 4-05,
>
> 2. Develop human performance metrics for efficiency, accuracy, and voter satisfaction,
>
> 3. Develop the performance benchmarks based on human performance data gathered from measuring current state-of-the-art technology, ..."

Also, Section 3.1 of the VVSG states that:

> "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."

This white paper explores possible high-level metrics in support of the development of these performance-based requirements. Note that no user-based pilot testing has yet been done, and that any performance metric must be supported by such testing before adoption.

2. Metrics: General Issues

There are certain properties that any good metric must possess.

- **Objectivity:** The metric must be derivable from plainly observable public facts. There should be no need for subjective judgment in the gathering of the data upon which the metric is based, nor in the computation of the metric from that data.
- **Intuitiveness:** The metric must correspond to common-sense notions of what it means for voting equipment to *accurately* and *quickly* capture voters' intentions. This means that there should be no obvious "counter-examples" - e.g. it should not be possible to construct a situation in which equipment A gets a lower accuracy score than equipment B, but yet would be intuitively judged as more accurate.
- **Appropriate Level of Detail:** All metrics reflect certain aspects but are insensitive to others. The metrics presented herein are, by design, *broad*, *bottom-line* measures. Accordingly, they are based on a “broad” task: filling out an entire ballot of moderate complexity (more details below, under Generalizability). They do not claim to diagnose particular problems with a voting system, nor analyze why a given system may be faster or more accurate than another. This approach is appropriate and typical for conformance testing and summative usability testing.
- **Technology-Independence:** Certain requirements in the VVSG pertain only to DREs or to paper-based systems and such type-specific requirements cannot be entirely avoided. Nonetheless, it is preferable that performance requirements and metrics apply equally to all types of systems to allow for fair comparison and evaluation. Except for some slight system dependence in the case of the speed metric, the metrics suggested herein all apply uniformly.
- **Repeatability:** Repeated measurements under the same conditions must yield reasonably consistent results. The metrics we propose are based broadly on the primary data and thus are not sensitive to small changes in that data. We hope that the primary data (e.g. vote totals and elapsed times, see Section 3) will be reasonably consistent from one set of subjects to the next, but of course, we cannot guarantee this until pilot testing has taken place.
- **Generalizability:** The measured test results must bear a reasonable correspondence to actual "real-world" results. This is perhaps the most difficult challenge for any performance-based voting metric, since we cannot directly measure the "true" accuracy of voting. It is not clear whether one could legally measure even the time taken to vote in an actual election. And even if it were possible to take such measurements in a "real-world" situation, there would be no way to control for demographic composition or ballot complexity.

  In order to make the testing more realistic, NIST has developed a database of actual ballots. Our initial test ballot is based upon an analysis of these actual ballots and its level of complexity is reasonably representative of what voters typically see. Finally, we believe that the metrics proposed are simple and direct enough that the *relative* performance of voting systems in a testing situation and in actual use would correspond.

- **Common Practice:** The performance metrics follow accepted practice for measuring effectiveness, efficiency, and satisfaction, as described in such documents as ISO 9241-11 [1] and the CIF [2].
- **Transparency:** The metrics have to be understandable and acceptable to the voting community**,** e.g., election officials, vendors, and testing laboratories.

3. Data Gathering

3.1 Accuracy

The current plan is to give subjects instructions on whom to vote for within an experimental "election" and then compare those dictated "intentions" to the actual votes cast. Ideally, we would have a record of every ballot cast by every subject. In the case of voting systems with paper ballots (or with VVPAT records), this data is available (although data collection might be labor-intensive). In the case of purely electronic systems, by design, it may be difficult to obtain individual electronic ballot information (although this is a question under study). This means that perhaps only total results can easily be measured. We propose to overcome this problem by separating subjects into groups which will be given identical instructions, and then inspecting totals for each group.

3.2 Speed

Obviously, any speed metric will depend on the actual time spent by each voter to complete the ballot. The main issue here is to have well-defined events that count as the "natural" beginning and end of the voting session. These events may be system-specific - for instance, in some systems, the voter uses a smart card to initiate the session. Other issues to be resolved are 1) whether external assistance is to be made available to subjects, 2) how to measure completion when the subject fails to cast a ballot, and 3) whether to impose a maximum time for a session. Pilot testing will be used to resolve these issues.

3.3 Satisfaction

The primary data in support of a satisfaction metric will be the results of a questionnaire distributed to the participants in the pilot testing. We are currently designing the survey instrument.

4. Accuracy

The proposed broad accuracy metric is based on the number of mistaken votes in relation to the number of votes that would be cast on a perfectly executed ballot. We define error rate as follows:

$$\text{error\_rate} = \frac{\#\text{mistakes}}{\#\text{votes on perfect ballot}}$$

We define a mistake as either 1) an omitted vote for a candidate for whom a vote was intended (negative mistake) or 2) a vote cast for a candidate for whom the vote was not intended (positive mistake). Error rate can be interpreted as the average number of mistakes a voter makes with each attempt to cast a vote. Note that the word “mistake” is not intended as a criticism of the voter, but simply to mean an omitted or miscast vote.

One of the attractive features of this simple definition is that we can "read off" mistakes directly from the recorded totals. As an example, assume a voting session involving 50 subjects using a ballot with 3 contests:

------ Aggregate Data ------	#negative mistakes	#positive mistakes	#correct votes
Contest #1: (vote for one)	1	1	49
should be / actual			
A1 50 / 49			
B1 0 / 0			
C1 0 / 1			
Contest #2: (vote for one)	9	2	41
should be / actual			
A2 0 / 0			
B2 50 / 41			
C2 0 / 2			
Contest #3: (vote for three)	5	3	145
should be / actual			
A3 50 / 47			
B3 0 / 0			
C3 0 / 1			
D3 50 / 48			
E3 50 / 50			
F3 0 / 2			
Totals 250	15	6	235

Summary: 21 mistakes; error rate = 21 / 250 = 8.4%

We examined a number of formulas, and believe that this approach is the most direct and intuitive and feasible way to quantify accuracy from a system perspective. Other more detailed usability metrics which preserve the individual performance data are also under

consideration and will be described in future reports when usability testing data is collected and analyzed.

5. Speed

Speed could be defined as the mean time taken by subjects to complete a voting session. This is the most natural metric and there seems to be no reason *a priori* to resort to a more complex formula. Also, it is the preferred CIF [2] method. Nonetheless, we shall examine both the measured mean and median in pilot testing to see whether there are considerations (such as repeatability) favoring the latter.

6. Satisfaction

Satisfaction is more difficult to quantify than speed or accuracy. First, the underlying data are participants' responses to a questionnaire, and so by definition subjective in origin (the coded responses themselves are, of course, objective data). Second, the way in which one should summarize these data is less intuitive.

Assuming that the questionnaire uses a Likert scale, or similar technique, the following is a possible approach. For each item (e.g. "Were you confident that your vote was recorded correctly?"), one could derive a summative metric for all the users, such as a mean or median. There could also be a "global" mean for all items. Then one could apply a separate benchmark to each item and also to the global response. Another approach might be to establish a "pass/fail" score for each item (e.g. 3 on a 5-point scale) and then determine the percentage of users who assign a passing score to the voting system. The benchmark would then be a minimum allowable percentage.

As with the other metrics discussed, the final shape of a satisfaction metric will depend strongly on the results of our pilot testing.

7. Conclusion

The TGDC and EAC have called for requirements for voting systems to include broad measures of effectiveness, efficiency, and satisfaction. NIST will be conducting user-based testing to determine if the metrics defined herein capture those qualities and if feasible and objective measurement methods to support them can be devised.

---

[1] ISO 9241 Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11 : Guidance on usability

[2] ANSI NCITS 354-2001 Common Industry Format for Usability Test Reports