Context Description: Posted Dec. 1, 2006

This draft report was prepared by NIST staff at the request of the Technical Guidelines Development Committee (TGDC) to serve as a point of discussion at the Dec. 4-5 meeting of the TGDC. Prepared in conjunction with members of a TGDC subcommittee, the report is a discussion draft and does not represent a consensus view or recommendation from either NIST or the TGDC. It reflects the conclusions of NIST research staff for purposes of discussion. The TGDC is an advisory group to the Election Assistance Commission, which produces voluntary voting system guidelines and was established by the Help America Vote Act. NIST serves as a technical advisor to the TGDC.

The NIST research and the draft report's conclusions are based on interviews and discussions with election officials, voting system vendors, computer scientists, and other experts in the field, as well as a literature search and the technical expertise of its authors. It is intended to help in developing guidelines for the next generation of electronic voting machine to ensure that these systems are as reliable, accurate, and secure as possible. Issues of certification or decertification of voting systems currently in place are outside the scope of this document and of the TGDC's deliberations.

On Accuracy Benchmarks, Metrics, and Test Methods

1 Preface

This document identifies problems with the test method for accuracy specified in VVSG'05 and describes some possible solutions. These possible solutions are the result of a preliminary analysis and do not constitute a NIST position or consensus. Harmonization with ongoing work on the similar topic of reliability testing has not yet occurred. We are providing this preliminary material only to keep the committee appraised of our activities and to provide opportunity for early feedback.

2 Accuracy test method

The informal concept of voting system accuracy is formalized using the ratio of the number of errors that occur to the volume of data processed, also known as error rate. By keeping track of the number of errors and the volume of data over the course of a test campaign, one can trivially calculate the observed cumulative error rate. However, the *observed* error rate is not necessarily a good indication of the *true* error rate. The *true* error rate describes the expected performance of the system in the field, but it cannot be observed in a test campaign of finite duration, using a finite-sized sample.

The system submitted for testing is assumed to be a representative sample (see [6] Ch. 8), so the variability of devices of the same type is out of scope. Valid concerns are the risk of rejecting a system whose true error rate is satisfactory ("producer's risk") and the risk of accepting a system whose true error rate is unsatisfactory ("consumer's risk").

2.1 Test design

VVSG'05 specifies a Probability Ratio Sequential Test in which the risk of accepting a system having a true error rate worse than 2×10^{-6} (1/500,000) and the risk of rejecting a system having a true error rate better than 10^{-7} (1/10,000,000) are both taken into consideration, yielding a test that specifies criteria for acceptance (i.e., accept if a given volume is achieved with less than r errors) as well as rejection. While it has numerous advantages, this test design leaves the test lab in a quandary if errors occur in other parts of the test campaign, e.g., after the criteria for acceptance in the accuracy test have been satisfied. There is no way to feed those other observations into the evaluation of accuracy.

The draft requirements at the bottom of this document specify rejection any time there is sufficient evidence to show that the probability that the system is conforming is less than 10 %, but they specify acceptance only after the test campaign is exhausted. The risk of accepting a system having a poor true error rate is mitigated by the length of the test campaign. If the system survives the entire test campaign without meeting the criteria for rejection, the error rate that was demonstrated with 90 % confidence is calculated from the collected data and recorded in the test report. In this way, the available data are put to maximum use and the test lab's quandary is avoided.

Both the VVSG'05 accuracy evaluation and the revised one rely on the simplifying assumption that probability of an error occurring is the same for each unit of volume processed. In reality, there are random errors that satisfy this assumption but there are also nonrandom errors that do not. For example, a logic error in tabulation software might be triggered every time that a particular voting option is used. Consequently, a test campaign that exercised that voting option early and often would be more likely to indicate rejection of the system based on accuracy than a test campaign that used different test cases or merely delayed execution of the problem test cases until the very end. However, since the Guidelines require absolute correctness of tabulation logic, the only undesirable outcome is the one in which the system containing the logic error is accepted. Other evaluations specified in the Guidelines, such as functional testing and logic verification, are better suited to detecting systems that produce nonrandom errors. Thus, when all specified evaluations are used together, the different test methods complement each other and the limitation of the accuracy test method with respect to nonrandom errors is not bothersome.

2.2 Fixed length versus sequential test plan

The draft text retains the practice of terminating the test campaign at the point where there is sufficient evidence to show with a specified level of confidence that the system fails to satisfy the accuracy requirement. However, an alternative is to use a fixed length test plan and postpone the assessment of accuracy until the test suite is exhausted. This decision is insignificant from the point of view of accuracy: If the system shows an error within the interval that would merit rejection using the sequential test criteria (i.e, the probability of non-compliance is greater than p), the probability that continued testing would collect sufficient evidence to "redeem" the error is negligible (i.e., less than $1-p$). However, a fixed length test plan may be desirable for other reasons, such as to provide the customer with the maximum amount of information that can be collected in one pass or to deliver a complete test report to the EAC. If we receive no guidance one way or the other, we will continue to specify termination of the test campaign as soon as there is sufficient evidence to show with a specified level of confidence that the system fails to satisfy the accuracy requirement.

2.3 Validity as a system test

The accuracy test specified in the 2002 VSS and VVSG'05 is not required to be an end-to-end test but may bypass portions of the system that would be exercised during an actual election ([5] II.1.8.2.3).

The use of text fixtures that bypass portions of the system may lower costs and/or increase convenience, but the validity of the resulting test is difficult to defend. If a discrepancy arose between the results reported by test labs and those found in state acceptance tests, it would likely be attributable to this practice.

The current draft Testing Standard of VVSG'07 has tightened requirements to prohibit bypassing portions of the voting system that would be exercised in an actual election. In the next section we discuss the ramifications this has for the accuracy benchmark.

3 Accuracy benchmark

The 2002 VSS and VVSG'05 contain requirements for the ballot position error rate to be no worse than 10^{-7}, but with the caveat that the error rate demonstrated by testing need be no better than 2×10^{-6} ([5] I.4.1.1, II.4.7.1.1). While both the upper and lower benchmark for the Probability Ratio Sequential Test are thus specified, two other critical parameters, the producer's risk (probability of rejecting a system that actually satisfies the more stringent benchmark) and the consumer's risk (probability of accepting a system that does not satisfy the more lax benchmark) are not specified. The test method sets both of these at 5 %.

If the testing is expected to provide a low consumer's risk under conditions simulating actual election use, the accuracy benchmark should be relaxed to what can practically be demonstrated in a testing process similar to the California Volume Reliability Testing Protocol [4]. Although we do not have complete information on the California volume tests, it appears likely that the total volume generated is significantly less than the volume required to achieve a consumer's risk of 5 % with a benchmark of 2×10^{-6}. (By the model specified in VVSG'05, this volume is 1,549,703; by the model specified in the draft text below, it would be 1,497,867.)

On the other hand, if a greater consumer's risk is acceptable, the higher benchmark can be retained. If a one-sided confidence interval is used, it is reasonable to specify a single benchmark of 10^{-7} with a producer's risk of 10 %. The resulting test protocol specifies rejection of a system that shows a single error with volume of less than 1,053,606. Intuitively, since the target error rate is so strict, any system that shows an error in the limited volume of testing that can practically be executed probably does not satisfy the requirement. If the producer's risk is left at 5 %, the single error cutoff instead falls at a volume of 512,933; i.e., if the first error happens at a volume between 512,933 and 1,053,606, we may be 90 % confident that the system is non-conforming but we cannot be 95 % confident.

The specific choice of benchmark and producer's risk may be the subject of lively debate, but adjusting these numbers is considerably easier than changing the test method.

4 Accuracy metric

The accuracy metric in the 2002 VSS and VVSG'05 is problematic in several ways.

a. Rather than a single, end-to-end system accuracy requirement, an error rate is applied separately to each low-level operation in the process (e.g., detecting selections on an optical scan ballot, storing selections into DRE memory, etc.) ([5] I.4.1.1). Most of these low-level operations are unobservable, hence the requirements are untestable. Moreover, there is no demonstrable relationship between the end-to-end error rate and the error rates of the low-level operations. Low-level errors may be amplified or corrected by other elements of the system.

b. The metric to be used in accuracy testing is ambiguous. The specified test protocol ([5] II.C.5) conflates ballot positions (voted or unvoted) with votes (only the voted ones) in the determination of volume. The test protocol in the 1990 VSS uses votes, not ballot positions ([2] F.5 and F.6).

c. The Bernoulli process assumed by the 2002 VSS and VVSG'05 is an invalid model of tabulation. A Bernoulli trial can only succeed or fail, hence the number of errors should be bounded by the number of ballot positions. In fact, a system can do worse than count every ballot position incorrectly: It can manufacture an unbounded number of additional "phantom votes" out of thin air. The Poisson process is a more valid model, allowing for the possibility of more than one error per unit of volume.

d. In the determination of error, it is unclear how inaccuracies in ballot counts and totals of undervotes and overvotes factor in. It is possible that the main vote total could be as expected but one of the other numbers could be completely wrong.

In the working draft, these problems have been remedied by defining a new metric, report total error rate, that considers only the accuracy of every count and total appearing in a vote data report. This expunges the untestable requirements on individual, low-level operations.

Because a single failure may now be amplified into more than one observed error (e.g., incorrect scanning of a contest could impact both the vote total and the overvote total), the accuracy benchmark may need adjustment to accomplish the intended level of accuracy, depending on what was intended. On the other hand, there is now an (equal and opposite?) opportunity for failures to be masked by "compensating errors." As stated in the previous section, adjusting the benchmark is much simpler than changing the test method.

5 Text from current CRT Working Draft

5.1 Product standard

→ 5.3.2-B End-to-end accuracy benchmark

All systems shall achieve a report total error rate of no more than 10^{-7} (1/10,000,000).

5.2 Testing standard

→ 5.3.2-A Calculation of report total error rate

Given a set of vote data reports resulting from the execution of test cases, the observed cumulative report total error rate shall be calculated as follows.

a. Define a "report item" as any one of the numeric values (totals or counts) that must appear in any of the vote data reports. Each ballot count (see Volume III Section 6.9.3.2) and each vote, overvote and undervote total for each candidate or measure (see Volume III Section 6.9.3.3) is a separate report item.

b. For each report item, compute the "report item error" as the absolute value of the difference between the correct value and the reported value. Special cases: If a value is reported that should not have appeared at all (spurious item), or if an item that should have appeared in the report does not (missing item), assess a report item error of one.

c. Compute the "report total error" as the sum of all of the report item errors from all of the reports.

d. Compute the "report total volume" as the sum of all of the *correct* values for all of the report items that are supposed to appear in the reports. Special cases: When the same logical contest appears multiple times, e.g. when results are reported for each ballot configuration and then combined or when reports are generated at multiple reporting levels, each manifestation of the logical contest is considered a separate contest with its own correct vote totals in this computation.

e. Compute the observed cumulative report total error rate as the ratio of the report total error to the report total volume. Special cases: If both values are zero, the report total error rate is zero. If the report total volume is zero but the report total error is not, the report total error rate is infinite.

Source: Revision of [2] F.6

→ 5.3.2-B Error rate data collection

During all test executions, the test lab shall keep track of the report total error and report total volume accumulated across all tests.

→ 5.3.2-C Error rate pass criteria

If a test case runs to completion, the test lab shall inspect the data reports and verify that counts and totals are reported in compliance with the requirements in Volume III Section 6.9. If all reported counts and totals are identical to the specified values, the test verdict shall be Pass. Otherwise, the following system-level accuracy decision criteria shall be applied:

a. If analysis of the cumulative behavior across all tests executed so far indicates that the probability of the true report total error rate being worse than the benchmark specified in Requirement III.5.3.2-B is greater than 90 %, the test verdict shall be Fail, the test campaign shall be terminated, and the system shall be rejected.

b. Otherwise, the error(s) and statistics shall be noted in the test report, the test verdict shall be assigned based on the other inputs (disregarding the error), and testing shall continue.

DISCUSSION

For a report total error *r* > 0, report total volume *t*, and end-to-end accuracy benchmark *l*, the probability that the true report total error rate is worse than the benchmark is equal to the probability that a system with true error rate *l* would show less than *r* errors under the same conditions, which is the value of the Poisson cumulative distribution function,

$$P(r-1; lt) = \sum_{x=0}^{r-1} \frac{e^{-lt}(lt)^x}{x!}$$

In Octave[1] version 2.1.73, this value can be calculated by entering poisson_cdf(*r*−1,*lt*).

If P(*r*−1,*lt*) = 0.9, the probability of the true report total error rate being worse than the benchmark is equal to, but not greater than, 90 %, so the test campaign is not terminated in this boundary case.

The report total volumes below which a given number of errors indicates rejection, for values less than 10, are shown in Table 5.

Solving for *t* in the trivial case, the volume below which a single error is grounds for rejection is given by

$$t_1 = \left\lceil \frac{-\ln(0.9)}{l} \right\rceil$$

Similarly, if a test campaign completes with zero errors after a volume of *t*, the error rate that was demonstrated with 90 % confidence is given by

$$l = \frac{-\ln(0.1)}{t}$$

Impact: Harmonized confidence level to 90 % (same as MTBF). A higher confidence level corresponds to a decreased chance of rejecting a bad system. Open question whether to truncate test campaign or allow it to run to completion.

Report total error	**Report total volume**
1	1053606
2	5318117
3	11020654
4	17447696
5	24325911
6	31518981
7	38947669
8	46561182
9	54324681

Table 5 Error rate cutoff points

6 Questions

a. Is the new approach OK overall? Default action: yes.

b. Is the 10^{-7} benchmark still appropriate? If not, what should it be? Default action: 10^{-7}.

c. Is the 90 % confidence level appropriate? If not, what should it be? Default action: 90 %.

d. Should the test plan be fixed length, or should we stop as soon as there is sufficient evidence that the accuracy benchmark is not satisfied? Or, should it be up to the vendor to decide if they want to pay for more testing even though they will probably fail anyway? Default action: stop as soon as there is sufficient evidence.

7 Nits

For a Probability Ratio Sequential Test, Epstein and Sobel [1] observed that truncation of testing results in the actual producer's risk being slightly different than the nominal producer's risk and invented an approximate correction factor. Said correction factor is used in [3], which in turn is used in [5]. Possibly an analogous correction factor is applicable to the proposed test design, though from the discussion in [1], it appears that the difference is negligible.

8 Bibliography

[1] Benjamin Epstein and Milton Sobel, "Sequential Life Tests in the Exponential Case," Annals of Mathematical Statistics v. 26 no. 1, 1955-03, pp. 82-93.

[2] Performance and Test Standards for Punchcard, Marksense, and Direct Recording Electronic Voting Systems, January 1990 edition with April 1990 revisions, in Voting System Standards, U.S. Government Printing Office, 1990.[2] Available at http://josephhall.org/fec_vss_1990_pdf/1990_VSS.pdf.

[3] MIL-HDBK-781A, Handbook for Reliability Test Methods, Plans, and Environments for Engineering, Development, Qualification, and Production, 1996-04-01.

[4] California Volume Reliability Testing Protocol rev. 2006-01-31, available from http://www.ss.ca.gov/elections/voting_systems/volume_test_protocol_final.pdf.

[5] 2005 Voluntary Voting System Guidelines, Version 1.0, 2006-03-06, available from http://www.eac.gov/vvsg_intro.htm.

[6] U.S. Election Assistance Commission, Testing and Certification Program Manual 2006, FR draft, 2006-09-28. Available at http://www.eac.gov/docs/Voting System Testing and Certification Program Manual FR DRAFT (Sept 28).pdf.

Notes

[1] Commercial equipment and materials are identified in order to describe certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

[2] The 1990 Voting System Standards package also included "A Plan for Implementing the FEC Voting System Standards," "System Escrow Plan for the Voting System Standards Program," and "A Process for Evaluating Independent Test Authorities."