Technical Guidelines Development Committee
May 21, 2007, Plenary Meeting

Benchmarks

David Flater
Computer Scientist

Terminology review: What is a benchmark?

- Definition: Quantitative point of reference to which the measured performance of a system or device may be compared
- Plain language: The number specified in the requirement (e.g., the failure rate shall not exceed [*benchmark*])
- The VVSG contains benchmarks for:
  - Reliability (failure rate)
  - Accuracy (error rate)
  - Rate of misfeeds for paper-based tabulators

Expectations

- Conformity assessment for these benchmarks targets "random" events (random failures, random errors, random misfeeds)
- It **may** collect nonrandom events (those traceable to design flaws and logic faults)
- However, such faults **should** be found first by other kinds of testing in the test campaign
  - Design review
  - Structural testing
  - Functional testing
  - Logic verification
  - Usability testing
  - Etc.
- No test is perfect—defense-in-depth

General Guidance from March Meeting

- Agree that old benchmarks deserve review and revision
- Plan "A" for determining new benchmarks not working—switch to plan "B"
- Get some "back of napkin" estimates of volume, tolerance for failures, etc.
- Use these to derive benchmarks that are in the correct order of magnitude
- Explain the reasoning ("show your work")

Specific Guidance from March Meeting

- Reliability: Any failure that results in even one ballot becoming unrecoverable (disenfranchisement) is unacceptable
- Accuracy: 1 in 10 000 000 benchmark considered arbitrary, possibly unattainable by paper-based systems

Terminology review: What is a failure?

- There is a precise (but complex) definition of failure designed more for arbitration than readability
- In plain language, failures are equipment breakdowns, including software crashes, such that continued use without service or replacement is worrisome to impossible
- Normal, routine occurrences like running out of paper are not considered failures
- Misfeeds of ballots into optical scanners are handled by a separate benchmark, so these are not included as failures

Reliability

- NASED representative provided estimates of volume, tolerance for failures, etc. for a medium-sized county in a western state
- Estimates were reviewed by other election officials
- Derived reliability benchmarks based on 1 % risk of exceeding tolerances
  - Special case: Benchmark for failures resulting in disenfranchisement set to zero
    - Falsifiable but not demonstrable—OK
- Explained the reasoning
  - Discussion backing up the estimates is preserved in the draft VVSG under Hardware and Software Performance, General Requirements, in Vol. III
  - Subsection explaining derivation of benchmarks using 1 % risk

Accuracy

- (Terminology review) Report total error rate—if the reported total is wrong, it's an error (or possibly several)
  - **Not** the human factors meaning of accuracy (usability testing)
  - Strictly a measure of mechanical performance
  - Bad inputs are thrown out
- Benchmark derived from the "maximum acceptable error rate" used as the lower test benchmark in VVSG 2005 (ballot position error rate of 1 / 500 000)
  - This was the rate that the test attempted to demonstrate
- Conversion from old metric (ballot position error rate) to new metric (report total error rate) explained in the discussion field of the requirement with more "back of napkin" reasoning

Misfeed Rate

- (Terminology review) Multiple feeds, misfeeds (jams), and rejections of ballots that meet all vendor specifications are all treated collectively as "misfeeds" for benchmarking purposes; i.e., only a single count is maintained
- Separate from reliability benchmark—Volume III, Requirements by Activity → Counting → Misfed Ballots
- Has ranged between 2 % (1 / 50) and 10^{-4} (1 / 10 000)
- Per input from NASED representative and election officials, now set at .002 (1 / 500)

Extra slide: Volume of testing

- VVSG'05 accuracy test required minimum of 1 549 703 ballot positions (possibly simulated volume on DREs)
- Op-scan
  - Volume test now specifies a minimum of 75 000 ballots (minimum value from 1990 VSS acceptance test guidelines)
  - Ballot style for testing is TBD by test suite; "back of napkin" estimates give 1 500 000 votes and 6 000 000 ballot positions
- DREs per California Volume Reliability Testing Protocol
  - Lower volume, but no longer simulated
- EBMs tested like DREs

Review of CRT Changes - I

Maintenance since previous meeting

- Productive discussions in teleconferences and e-mail
- Most changes were to clarify previously written requirements and definitions without changing their intent—such changes are numerous and not detailed here
- Procedural "requirements" changed to informative assumptions

Recent substantive changes

- Conformance clause
  - Added classes for activation device, audit device, CCOS
  - Brought back system-level classes for IDV, Election Verification (by request of STS)
  - New [STS] subsection about innovation class submissions
- Durability of paper: point to Government Paper Specification Standards

Summary of major changes since VVSG'05

- Refocused Terminology Standard (glossary) to provide well-formed terminology for the VVSG
- Separated documentation requirements (data to be provided) from functional requirements (product standard)
- Defined voting variations, system and device classes
- Identified requirements
- Specified applicability of requirements
- Revised benchmarks and related test methods
- Refocused coding conventions on integrity and transparency
- Defined COTS-related concepts better
- Clarified and strengthened optical scanning requirements
- Clarified reporting requirements
- Added logic model (definitions) and logic verification
- Added volume test
- Made consistent with current law, policy, and technology
- Removed redundant and problematic requirements

Review of CRT Changes - II

Alan Goldfine
Computer Scientist

Quality Assurance/Configuration Management

- Response to
  - TGDC Resolution 30-05
  - Statement of direction at December 2006 TGDC plenary that ISO 9000/9001 standard should provide the framework for new VVSG requirements dealing with quality assurance

Quality Assurance/Configuration Management
Changes from 2005 VVSG

- 2005 VVSG
  - Volume I: Sections 8 and 9
  - Volume II: Section 7
- Replaced by new VVSG
  - Volume 3: Section 16.4.2
  - Volume 4: Chapter 2
  - Volume 5: Section 4.4

Quality Assurance / Configuration Management
Changes since last plenary

- Revised the requirement dealing with the timing of the vendor delivery of the Quality Manual, per instructions from the TGDC
- Based on CRT comments, clarified and sharpened the informative text
- Incorporated the requirements into the draft VVSG

Electromagnetic Compatibility

- Goal:
  - To update the 2005 VVSG requirements to
    - reflect the latest available information
    - reference applicable standards, rather than repeating or excerpting text from those standards
    - clearly separate requirements from testing specifications

Electromagnetic Compatibility

Changes from 2005 VVSG

- 2005 VVSG
  - Volume I: Sections 4.1.2.4 - 4.1.2.12 and part of Section 6 (Telecommunications)
  - Volume II: Section 4.8
- Replaced by new VVSG
  - Volume 3: Sections 16.3.3 - 16.3.5
  - Volume 5: Sections 5.1.1 - 5.1.3

Electromagnetic Compatibility

Changes since last plenary

- Completed the requirements in all three categories:
  - Conducted immunity
  - Radiated immunity
  - Telecommunications immunity
- Discussed requirements at CRT meetings
- Made final edits to the informative text
- Incorporated the requirements into the draft VVSG