Usability Performance Benchmarks
For the Voluntary Voting System Guidelines

Prepared at the direction of the HFP Subcommittee of the TGDC

June 26, 2007
DRAFT

This paper has been prepared by the National Institute of Standards and Technology at the direction of the HFP subcommittee of the Technical Guidelines Development Committee (TGDC). It may represent preliminary research findings and does not necessarily represent any policy positions of NIST or the TGDC.

The Technical Guidelines Development Committee is an advisory group to the Election Assistance Commission (EAC), which produces Voluntary Voting System Guidelines (VVSG). Both the TGDC and EAC were established by the Help America Vote Act of 2002. NIST serves as a technical adviser to the TGDC.

.

1. Introduction

In the Help America Vote Act (HAVA) of 2002, the Election Assistance Commission, in consultation with the Director of the National Institute of Standards and Technology, was mandated to submit a report on human factors, usability, and accessibility to Congress. This report included two recommendations for the development of usability performance benchmarks for voting systems:

- Develop voting system standards for usability that are performance-based, high-level (i.e., relatively independent of the technology), and specific (i.e., precise).
- Develop a valid, reliable process for usability conformance testing of voting products against the standards described in the recommendation above with agreed upon usability pass/fail requirements.

The Human Factors and Privacy Subcommittee of the Technical Guidelines Development Committee, formed under HAVA, subsequently requested that NIST develop usability performance requirements for inclusion in the Voluntary Voting Systems Guidelines in Resolution #5-05 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. This conclusion is based, in part, on the analysis in the NIST Report, *Improving the Usability and Accessibility of Voting Systems and Products* (NIST Special Publication 500-256).
>
> Performance requirements should be preferred over design requirements. They should focus on the performance of the interface or interaction, rather than on the implementation details. … Conformance tests for performance requirements should be based on human performance tests conducted with human voters as the test participants. The TGDC also recognizes that this is a new approach to the development of usability standards for voting systems and will require some research to develop the human performance benchmarks and the test protocols. 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,
> 4. Develop a conformance test protocol for usability measurement of the benchmarks,
> 5. Validate the test protocol, and
> 6. Document test protocol."

This report summarizes the research conducted to develop the test, metrics, and benchmarks in response to this resolution and describes the resulting performance-based requirements included in the current draft of the VVSG. Supporting materials such as the test data and test materials can be found at

http://vote.nist.gov/PublicHearingsandMeetings.html under Supporting Materials for Usability Performance Benchmarks.

This research included:

- Defining a user-based test for measuring effectiveness, efficiency, and satisfaction of voting systems
- Defining the metrics by which the systems tested will be measured
- Validating the test methodology
- Determining that the test protocol is repeatable
- Setting performance benchmarks based on data collected by running the test on typical voting systems.

2. Defining the test

We developed a test for measuring effectiveness, efficiency, and satisfaction of voting systems. There were two challenges in defining this test.

1. The test itself had to be completely objective so that staff from an accredited Voting System Test Laboratory (VSTL) could conduct the test, analyze the data, and obtain results that are reliable (that is, repeatable under exactly the same conditions and reproducible by another lab).
2. We needed to determine the number of test participants sufficient to allow for statistical analysis and to ensure repeatability of the test on the same voting system.

We call this test protocol the NIST **Voting Performance Protocol (VPP).** The NIST VPP is intended to be a *repeatable, controlled experiment.* The goal is not primarily to provide a "realistic" measure of voting behavior, but rather to isolate and measure the effect of various voting systems – and this requires that all other potential variables be held as constant as possible. See Section 8 for an overview of the VPP.

2.1 The ballot and instructions

After conducting research into typical ballots from around the country, we designed a logical definition for the test ballot and the specific voting instructions. The test ballot included the following:

- Twenty contests, referenda, and ballot initiatives
- Contests at various depths of the government hierarchy (federal, state, local)
- Contests that were both partisan and nonpartisan
- Contests that were single member and multimember
- Retention races
- Constitutional amendments
- Ballot measures

To avoid bias, realistic but not real names were chosen for all candidates, fictitious and neutral party names were created using color names and fictitious but realistic constitutional amendments and ballot measures.

The test participants were told to make 28 specific voting choices for the test ballot. Various tasks were included in the instructions for the participants to represent realistic voting events. These included:

- Voting for names that appeared at various locations within a list of names
- Voting only a partial slate in a multimember contest
- Skipping specific elements of the ballot
- Write-in votes

By instructing the participants how to vote, we could determine difference between the intended votes of the test participants and the actual votes that they cast. We then measure the accuracy of the cast ballot by counting the number of correct votes, 28 being a perfect score.

2.2 The test participants

For conformance testing, we needed enough participants to provide reliable results that can be statistically analyzed, but not so many participants that conducting the test becomes financially infeasible. The number of participants needed is primarily determined by our need to calculate reasonably tight 95% confidence intervals. (We explain these calculations later in the paper.) We found that a minimum of 100 participants is required to allow for this statistical analysis.

Another important consideration was that the subject pool had to be consistent enough to allow repeatability. Variability of the subject pool represents “noise” which masks the “signal” we are trying to detect: the effect of the voting system. And finally, we wanted to use citizens who were reasonably representative of the actual voting population (See the Supporting Materials for Usability Performance Benchmarks for the detailed demographics).

2.3 The test environment and voting equipment

As these are controlled experiments, they were performed in a usability test lab so that the environmental conditions such as lighting could be held constant. We plan to specify this (and other steps) as comprehensively as possible in the next version of the VPP so that a test lab can duplicate the environment.

All the voting systems used in these tests were systems certified under the Voting System Standards 2002 (VSS 02) and used in at least one election. We asked the vendors to implement the standard test ballot for their systems to best showcase the usability of their equipment.

3. Defining how the systems will be measured

We defined the performance metrics for effectiveness, efficiency, and satisfaction as follows:

1. Effectiveness is measured as
   a. **Ballot Casting Rate** -- Percentage of users who successfully cast a ballot (whether or not the ballot contains erroneous votes). Failure to cast a ballot might involve problems such as a voter simply "giving up" during the voting session because of an inability to operate the system, or a mistaken belief that one has successfully operated the casting mechanism.

   b. **Correct Ballot Rate** – Proportion of cast ballots containing no erroneous votes. If a cast ballot contains even one mistaken vote (either a vote for an unintended choice, or a missing vote), then it is deemed incorrect.

   c. **Accuracy Index**[1] – A measure of overall voting accuracy. Each voter is given 28 "voting opportunities" within the ballot. The number of these that are correctly performed yields that voter's simple *accuracy score*. The higher the rate, the higher the resulting accuracy index for that voter. We calculate the accuracy index as a **capability index (C_{pl})** :

$$C_{pl} = \frac{\bar{X} - \mathrm{LSL}}{3S}$$

   where LSL is a lower specification limit, X-bar is the mean, and S is the standard deviation. We use 85% as our LSL. A complete description of process capability and how to use and calculate capability indices can be found in http://www.itl.nist.gov/div898/handbook/pmc/section1/pmc16.htm

2. Efficiency is measured as **Average Voting Session Time** – mean time taken per voter to complete the process of activating, filling out, and casting the ballot.

3. Satisfaction is measured as **Average Voter Confidence** – mean confidence level expressed by the voters that the system successfully recorded their votes. During our research to validate the test methodology we also measured satisfaction by administering a modified version of the Survey of User Satisfaction (SUS). (SUS is an industry standard questionnaire measuring overall satisfaction with a system.) Confidence is a measure of the participant's assessment of their own performance on the test. This is based on a confidence question we specifically developed for this test.

[1] The capability (accuracy index) is calculated based on the mean, standard deviation, and the selected lower standard level. Its approximate 95 % confidence interval is a mathematical derivation dependent on the calculated Cpl and the N, the sample size when N is large.

This paper has been prepared at the direction of the HFP subcommittee. It does not necessarily represent any policy positions of NIST or the TGDC.

For those wishing to read more about measuring usability and related statistical methods, we recommend starting with the site http://www.measuringusability.com .

4. Validating our test methodology

The first step in validating the test methodology was to show that we can detect performance differences between voting systems with our test and that these differences are what we expected. For example, we had two systems (we will call them A and B), which prior research had shown would be expected to have different performance on at least some of our measures. Twenty four (24) participants were asked to use System A and twenty- three (23) participants were asked to use System B using the test ballot and procedures we had defined.

Data collected from this test produced the following results for correct ballot rate, the system's accuracy index, voting time, and satisfaction using a questionnaire we developed. (We did not collect data for ballot casting rate.)

	Ballot Casting Rate **Confidence Intervals**	**Correct Ballot Rate** **Confidence Intervals**	**Accuracy Index** **Confidence Intervals**	**Voting Time (secs)** **Mean and Standard Deviation**	**Satisfaction** **Mean and Standard Deviation**
System A	N/A	42.6 -78.9%	.85-1.62	402.56 +/- 103.08	80.7% +/- 15.2%
System B	N/A	22.1 -59.3%	0-.24	431.65 +/- 114.30	73.6% +/- 16.3%

Table 1: Summary Performance Data by System for Validity Test

4.1 How we analyzed the test results

We used statistical analyses (a binomial calculation called Adjusted Wald) for the correct ballot rate confidence intervals. For the accuracy index, with a large enough sample size[2]; we can assume an asymptotic normal distribution, to calculate the 95% confidence intervals. These ranges are very likely to contain the true accuracy index of the systems. (We calculate a 95% confidence interval based on our sample of test participants. There is a well-defined true value (but unknown), e.g., correct ballot rate. We have a 95% confidence that for this one sample, this true value is contained in this fixed confidence

[2]It is recommended for a Cpl such as the accuracy index to have 100 participants to be able to assume a normal distribution for the purposes of measuring as system for conformance to a benchmark.

interval. We can say we are 95% confident that the calculated ranges contain the true value of the system for our measures.) These calculations showed us that we *are* able to detect differences in effectiveness between the two systems.

For correct ballot rate, of the 24 participants who attempted to vote using System A, 15 were able to cast the ballot without any deviation from the results expected (62.5%) and, of the 23 participants using System B, 9 were able to cast the ballot correctly (39%). The 95% confidence interval for System A is between 42.6% and 78.9% and for System B between 22.1% and 59.3%. Though there is overlap between these two ranges, this is likely due to the small number of participants in this evaluation. The data was in the direction expected, with System A showing higher effectiveness.

For the Accuracy Index, assuming a Lower Specification Limit of 85%, System A was measured at 1.24 and System B at 0.10. Calculating a 95% confidence interval for the accuracy index gives us a range of .85 to 1.62 for the System A, and 0 to .24 for System B. Thus, the accuracy index also showed a statistically significant difference in the effectiveness between the two machines.

As an additional check, a Mann-Whitney test was used to check for statistically significant differences between the accuracy scores of the two systems.[3] Using this calculation, a statistically significant difference (p-value>0.05) was seen, with System A performing better than System B.

A statistical t-test was used to compare time data between the two systems. This analysis of the data showed no statistically significant difference between the two systems though System A had a slighter faster time than the System B.

A t-test was also used to compare the satisfaction data collected from the satisfaction questionnaire. Again, there was no statistically significant difference in the data though the satisfaction data showed somewhat better scores for System A than for System B.

The results of this testing indicated that the test was valid, in that we were able to “tell the difference” between the two systems and it was what we expected. In this initial test, it was only effectiveness that could be used to discriminate between these two. However, as we shall see, the later benchmark tests, performed on a wider variety of systems, showed significant differences in efficiency. The satisfaction questionnaire didn't show statistically significant differences in the validation or any of the repeatability tests so we are not using it as a benchmark and we simplified it to a single question about confidence.

So far, none of our tests have shown strong correlations among the three types of metrics, e.g., effectiveness and efficiency do not appear to be strongly linked. Faster time does not necessarily imply accuracy or vice versa. One cannot expect that the significance (if any)

[3] A Mann-Whitney test, a nonparametric statistical analysis approach, was used because the performance data was not normally distributed.

This paper has been prepared at the direction of the HFP subcommittee. It does not necessarily represent any policy positions of NIST or the TGDC.

of the relative independence of the types of metrics is unclear, but does not call into question the test's ability to measure the various usability aspects.

5. Determining whether the test protocol is repeatable

In addition to test validity, we needed to demonstrate that the test was repeatable – that the test returned results with no statistically significant difference each time it was used by the same testers on the *same system*. Only the set of participants changed. Four tests of the same system were conducted. Between 44 and 50 individuals participated in each of these four tests.

To guarantee the independence of the data, the test was conducted the same way with each participant. All instructions about the test process were given to the participants in writing, with no additional individual assistance offered. The following statement was read by the test administrator in response to any question from a participant:

> "I'm sorry but I'm not allowed to help you or answer any questions. If you are having difficulties, you can try to finish. When you feel you are finished, you may just leave the machine."

Data collected from this test produced the following results as shown in Table 2 for cast ballot rate, correct ballot rate, the system's accuracy index, and voting time:

Repeatability Test	Ballot Casting Rate Confidence Intervals	Correct Ballot Rate Confidence Intervals	Accuracy Index Confidence Intervals	Voting Time (secs) Mean and Standard Deviation)
Test 1	87.4-100%	53.4-80.1%	.11-.33	695.4+/-224.0
Test 2	88.3-100%	40.3-67.4%	.05-.24	662.0 +/-245.0
Test 3	77.8-96.0%	31.7-59.9%	.05-.25	691.0+/-206.8
Test 4	86.3-99.7%	44.0-70.6%	.19-.41	633.8 +/-166.5

Table 2: Summary Repeatability Performance Data by Test

- Ballot casting rates: For these four tests, 44 of 45, 48 of 49, 44 of 49, and 50 of 52 participants finished voting. The confidence intervals overlap indicating that the data are repeatable.

- Correct ballot rates: Participants who were able to complete voting were the smaller number in the voter casting rate data (just above). For these four tests, the number of correct ballots as compared to cast ballots was 30 of 44, 26 of 48, 20 of 44, and 29 of 50. Again, the confidence interval ranges overlap indicating that the data are repeatable.

- The accuracy index confidence intervals for these four tests show sufficient overlap to indicate that these are consistent results

- The voting time was also analyzed and shown to be consistent for all tests (p-value >0.05).

The Mann-Whitney test was again used to determine if the accuracy scores of the four tests were consistent and they were (p-value > 0.05) except for Test 1 compared to the Test 3 – and these two tests were only marginally different.

6. Setting the benchmarks

Four different voting systems were tested to determine possible values for benchmarks for each of the three effectiveness measures. The systems were from four different vendors and representative of current systems and technology. Four sets of individuals were recruited, one for each of the four systems, using the demographic criteria previously defined. Out of 52 participants, System A had 50 that voted successfully; System B had 42 out of 42, System C had 42 out of 42, and System D had 47 out of 50.

Data collected from this test produced the results in Table 3 for ballot casting rate, correct ballot rate, accuracy score, and the system's accuracy index:

	Ballot Casting Rate **Confidence Intervals)**	**Correct Ballot Rate** **Confidence Intervals)**	**Accuracy Score and Standard Deviation** **Mean and Standard Deviation**	**Accuracy Index** **Confidence Intervals)**
System A	86.3-99.7%	44.2-70.6%	95.0 +/- 11%	.19-.41
System B	92.8-100%	42.2-70.9%	96.0 +/- 6%	.49-.85
System C	92.9-100%	22.4-49.9%	92.4 +/- 13.0%	.08-.30
System D	83.2-98.6%	50.4-76.6%	92.4 +/- 19.0%	.03-.22

Table 3: Summary Benchmark Performance Data by Voting System

Data collected from this test produced the results in Table 4 for voting time and confidence:

	Voting Time (secs) Mean and Standard Deviation	Confidence Mean and Standard Deviation
System A	638.1 +/- 166.1	4.0 +/- 1.0
System B	429.3 +/- 156.3	3.3 +/- 1.4
System C	870.7 +/- 236.0	3.6 +/- 1.4
System D	744.7 +/- 209.3	3.8 +/- 1.2

Table 4: Benchmark Time and Confidence Data

Based on the data collected in these tests, we can derive benchmarks for requirements that must be met in conformance tests for the VVSG. **It is important to note that we chose these benchmarks so that they are achievable by some systems, but not so low as to be trivially easy to meet with any of the current voting system implementations.**

6.1 Confidence Intervals

The conformance tests are designed to give the tested system the benefit of the doubt. By definition, we are 95% confident that the "true" value of the system for a given metric lies within the confidence interval (CI) resulting from a test. Therefore, if the entire CI falls below the benchmark, we are at least 95% confident that the system has not attained the benchmark.

6.2 Ballot Casting Rate

The ballot casting rate for all the systems tested was relatively high and showed that is it possible for a system to show a ballot casting rate of nearly 100%. It seems reasonable to set the value for this rate at a fairly high level for the conformance test since this value represents all but the few individuals who could not finish once they started the voting task

The ranges of values obtained in this set of tests suggest setting the benchmark for the Ballot Casting Rate at 98%.

As an example of what this means, if there were 100 participants who attempted to use a system, that system would pass the benchmark if at least 95 participants succeeded, because the resulting 95% confidence interval of [.8854, .9813] include the 98%

benchmark. If more than five participants failed to complete the test, then the system would fail on this conformance measure.[4]

6.2 Correct Ballot Rate

The ability for the participants to vote a ballot with no errors was lower than might be expected in all systems tested. We can suggest two reasons for this finding: first, the test itself represents a moderately difficult task in which participants are required to match values from the instructions for 20 separate contests. Second, this rate represents any error out of the 28 specific votes (or non-votes) expected. Many participants had only a single error on their ballots.

The ranges of values obtained in this set of tests suggest setting the benchmark for the Correct Ballot Rate at 70%.

The calculation of the confidence interval is the same as for the ballot casting rate above.

This value would allow the majority of the systems tested to pass the conformance test; though retesting these machines with 100 participants would decrease the confidence interval and therefore make the test stricter.

6.3 Accuracy Index

The accuracy index levels in this test showed a wide range of results, with two of the systems significantly below the other two systems. This difference in accuracy index was not the result of significant differences in the raw accuracy score (which was in the 90% range for all four systems), but in the large differences in the standard deviations. A high standard deviation reflects high variability among the performance of individual voters. A lower deviation implies a more equal level of performance among participants. The mean and standard deviation for each of the machines tested are shown in Table 3 above.

We calculated a conformance level for the accuracy index with a goal of a mean of at least 92.5% and a standard deviation of no more than 7%.[5] Using a lower standard limit of 85%, the calculated cut off point for the accuracy index is .35, and this is our recommended benchmark. Hence,

The data suggests setting the benchmark for the Accuracy Index at .35.

This value allows two of the systems tested to fail. However, our observation of participants during testing and an analysis of the nature and type of errors indicate that improvements (and better results) could be obtained in all of these designs.

---

[4] The reader can perform the calculations for this example at http://www.measuringusability.com/wald.htm using the calculator and the Adjusted Wald confidence interval.

[5] It should be noted that, since the accuracy index combines the mean and standard deviation, is not necessary for a machine to meet these individual requirements provided the calculated accuracy index is acceptable.
This paper has been prepared at the direction of the HFP subcommittee. It does not necessarily represent any policy positions of NIST or the TGDC.

6.4 Voting time, satisfaction, and confidence

In analyzing the research data, time did not correlate with effectiveness and does not necessarily imply that a system is not usable. It may be possible to use a simple time benchmark to eliminate very slow systems but there is a large standard deviation.

We tested satisfaction in two ways: in earlier tests, we used a 10-point questionnaire, and in later tests we asked only two questions about the participants' confidence that they had voted as instructed. Neither measurement appeared to be sensitive enough to be used for a performance benchmark.

This is the revised confidence question that we used during Benchmark testing.

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
	1	**2**	**3**	**4**	**5**
If I had to use this machine in a real election, I'd be confident in my ability to enter my votes and I'd trust this machine.					

The TGDC has decided not to make time or confidence pass/fail requirements in the VVSG. We recommend that Average Voting Session Time and Average Voter Confidence be tested and reported by the Voting System Test Laboratories (VSTLs) because election officials may find them helpful for determining numbers of vote stations needed and general attitudes held by voters towards the systems.

6.5 Recommended number of test participants

Based on the above experiments, we recommend 100 test participants. At that number the 95% confidence intervals are small which means we are getting results very close to the true performance of a voting system with a high degree of confidence. This number is also recommended in using the Cpl-type measure of the accuracy index so that a normal distribution can be assumed. The cost to recruit 100 participants and run the test is reasonable and similar to the typical cost for testing by a VSTL.

7 The performance requirements in the VVSG

The June 21, 2007 version of the VVSG usability performance requirements are based on these results. They are:

3.2.1.1-A Ballot Casting Performance

The system shall achieve a ballot casting rate of at least 98% as measured by the NIST VPP.

3.2.1.1-A Correct Ballot Performance
The system shall achieve a correct ballot rate of at least 70% as measured by the NIST VPP.

3.2.1.1-B Accuracy Index Performance
The system shall achieve an accuracy index of at least 0.35 as measured by the NIST VPP.

Reporting requirements in VVSG, Volume IV include reporting the above results and:

Voting Session Time
The test lab shall report the average voting session time, as measured by the NIST VPP.

Average Voter Confidence
The test lab shall report the average voter confidence, as measured by the NIST VPP.

8 Voter Performance Protocol (VPP)

This is an overview of the VPP we used. Eventually intended for the VSTLs, this user test protocol will be described in much greater detail as the test methods in support of the VVSG are developed.

1. **Ballot onto voting system.** The voting system vendor is responsible for putting the test ballot onto the system to be tested.

2. **Set up.** Two machines are set up for each test: the machine being tested and a machine whose value on the test had been previously established. This is a way to calibrate the lab testing process. For the validity and repeatability research described in this white paper, it was not necessary to use a validation machine, as we were not certifying systems to conform to the benchmarks, but proving the validity of our method and generating the benchmarks.

3. **Participants and their demographics.** Participants are recruited for the test based on specific demographic characteristics of age, education, race, and gender, and location.

4. **Test environment.** The voting systems are installed in a lab under realistic but controlled conditions that make it easy for participant use (e.g., presence of a chair, appropriate lighting).

5. **Method.**

   a. Participants arrive at the test lab, fill out the consent form, and wait in a room separate from the voting system until it is their turn to participate. They are provided with the overview of the test procedures (i.e., that they would be asked to enter specific votes provided to them, and that the test facilitator would not be able to assist them once they started)
   b. When it is their turn to participate, a test facilitator shows them the voting system to which they are assigned and gives them the full voting instructions. The participants are also given any additional materials needed to use the specific voting system (e.g., access code, paper ballot, voting card) but they are not told how to use them.
   c. To minimize interference in the measurement of usability, once the participant has begun the test, the facilitator's interaction with them is limited to the following statement: **"**I'm sorry but I'm not allowed to help you once you start. If you are having difficulties you can try to finish.. If you are stuck and cannot continue, you can stop if you wish."
   d. Behind a mirrored glass, observers time the participants as they complete the voting task.
   e. After the participant completes casting vote or stops, the facilitator directs the participant to a third room where another individual administers the post-test questions and compensates them for their participation.

6. **Data handling.** A tally is kept of all participants who fail to complete the voting task either by leaving prior to completion or by leaving believing they had completed the task but had not.. For all participants that complete the voting task, the ballot data as counted by the system is entered into an Excel spreadsheet where they are tallied to determine the numbers of correct votes and the number of ballots cast without any errors. From these data, a comparison is made between the values obtained from the test validation machine and from previous values obtained for the same machine. If the comparison shows the data to be consistent, the test is considered valid and analysis is performed on the machine being tested. The data from machine being tested is converted into 95% confidence intervals for the three measures of interest (ballot casting rate, correct ballot rate, and accuracy index). Additional data, such as timing, post test questions, and demographic information, is entered into separate Excel spreadsheets for reporting.

9 Next steps

The research described in this white paper shows the validity and repeatability of our testing methodology and allows us to determine benchmarks and thus specify usability performance requirements for voting systems. This methodology is technology-independent. Any type of voting system that can implement the test ballot can be evaluated for conformance to these requirements. However, to ensure that this

methodology can be reproduced in any VSTL, we will be performing research to determine how precisely the test conditions must be followed as part of our work on VVSG test method development. If one tries to replicate what we did with the same recruiting criteria and same "conditions" than one should get the same results. But "conditions" in our case would include the same database we recruited from and the same exact demographics. We need to determine how we can repeat the test in other parts of the country to generate the same results. We do not know at this point if there are regional differences that produce different results

This paper has been prepared at the direction of the HFP subcommittee. It does not necessarily represent any policy positions of NIST or the TGDC.