USER-CENTERED DESIGN
20548 Deerwatch Place, Ashburn, VA 20147 TEL: (703) 729-0998 www.user-centereddesign.com

Residential Climate Control Units: Usability Test Validation

July 2011

Submitted to:
Navigant Consulting, Inc.

Table of Contents

1) Introduction

User-Centered Design, Inc. was contracted by Navigant Consulting, Inc. to perform a review and assessment of a user-based, performance test developed for testing the usability of Residential Climate Control Units. The test consisted of 4 tasks selected from a larger set of tasks used in testing. The four tests selected for the specific unit under test (UUT) were:

Task 1: Set the unit to the correct time.
Task 2: Read the current ambient temperature and the currently active setpoint.
Task 3: Set the unit to control heating
Task 4: Program the unit to control the temperature throughout part of a single day of the week.

A total of 50 participants were recruited to perform the tests. The participants were selected to meet demographic criteria specified in the proposed test method. The test method required specific ratios for gender, age, education and colorblindness. Data was collected on pass/fail performance and total time on task. Participants were also provided with a post-test survey that asked them about their prior experience with programmable thermostats and their general knowledge of HVAC systems.

The data from testing was analyzed for the effect of sample size on the results by selecting subgroups of the total population at various sample sizes including 7, 14, and 28 participants per sample. Each sample was selected to ensure that the demographics of the selected sample were equivalent across each sample size. In addition, the samples were selected to ensure the least number of repeated data points between each sample. Pass/fail performance data, total time on task, and the Time & Success metric described in the test method were calculated for each subgroup. The results of the survey were also compared to the total group to determine if there was an interaction effect between prior knowledge and/or knowledge of HVAC systems with performance.

2) Data Analysis

The main question posed in this research was the effect of sample size on test reliability. To address this question, the total population of 50 data points was resampled to create simulated populations with sizes equal to 7, 14, and 28. Each sample was created to maintain the original demographic ratios. To ensure that the effect of sample size on test reliability was being evaluated, samples drawn from the larger populations were selected with the fewest number of repeated data points. Also generated were the two most extreme cases based on the data (one data set with the lowest possible set of values and one with the highest possible set of values—possible within the larger sample). The task completion rate, time-on-task data, and proposed Time & Success metric were analyzed separately.

a) Time on Task Data

Prior to calculating mean time-on-task, the time data was tested to determine if it was normally distributed. Based on the normal quantum tile plot method, only the time-on-task data for Task 1 was shown to be normally distributed in its natural form. The data for Task 2 and Task 3 was then transformed using a natural log function, which has been shown in some cases to produce normally distributed data for time-on-task data. The resultant analysis of the transformed data for Task 2 and Task 3 was then tested; it was also shown to be non-normally distributed. It is assumed that interface design accounted for the variance in distribution for the time data for the different tasks. However, since these differences appear to be dependent on the interface design and not the task, it is not possible to predict whether the resulting data for any specific design would be normal.

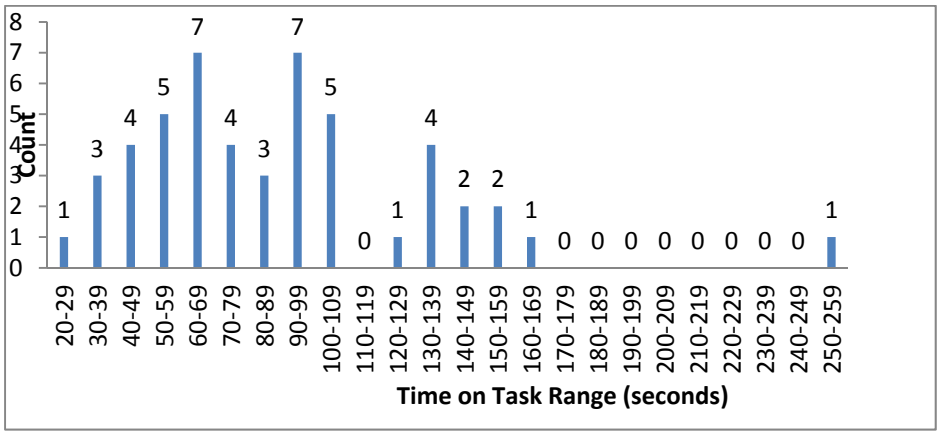

Figure 1: A Histogram of the Task 1 Time on Task Data

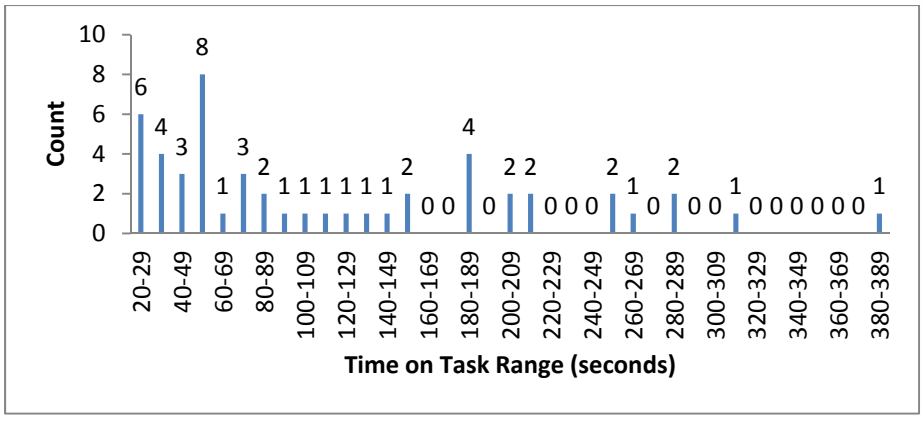

Figure 2: A Histogram of the Task 2 Time on Task Data

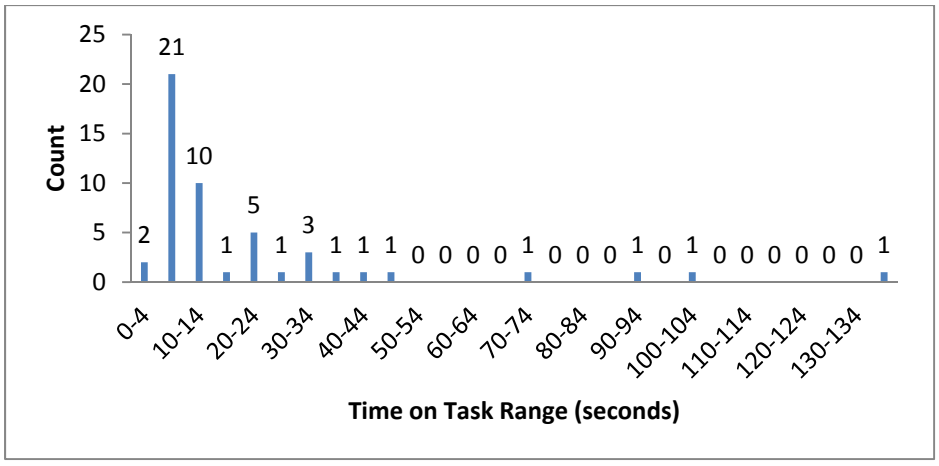

Figure 3: A Histogram of the Task 3 Time on Task Data

Performance data was generated using the binomial calculation based on task success rate. This formula calculates a predicted value for performance and an associated confidence interval based on the sample size and actual completion rates.

Resampling analysis was performed on Task 1, 2, and 3 using Task Success Rate data. (Poor performance on Task 4—with only two individuals being successful—was the reason it was removed from these analyses.) The tables below show the pass/fail performance results, by task, for the resampled groups, by task number, at the various sample rates (to create groups of 7, 14, and 28).

Sample	Task 1	Task 2	Task 3
1	100%	43%	71%
2	85%	71%	71%
3	85%	100%	71%
4 Lowest Possible	57%	0%	29%
5 Highest Possible	100%	100%	100%

Table 1: Predicted Values for Completion Rate for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

Sample	Task 1	Task 2	Task 3
1	86%	29%	79%
2	86%	71%	64%
3	76%	50%	50%
4 Lowest Possible	64%	21%	43%
5 Highest Possible	100%	86%	93%

Table 2: Predicted Values for Completion Rate for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

Sample	Task 1	Task 2	Task 3
1	82%	43%	71%
2	83%	64%	54%
3 Lowest Possible	67%	39%	46%
4 Highest Possible	92%	71%	86%

Table 3: Predicted Values for Completion Rate for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

It can be seen from these tables that there is a wide variation in completion rates. This variation can also be seen to be inversely correlated with the sample size (e.g., there is more variability in smaller sample sizes) with variances as high as 100% for samples with 7 participants to 43% or less for samples with 28 participants.

The 95% Confidence Interval was also calculated for each of the resampled groups plus the lowest and highest possible values. These values are plotted in the figures below.

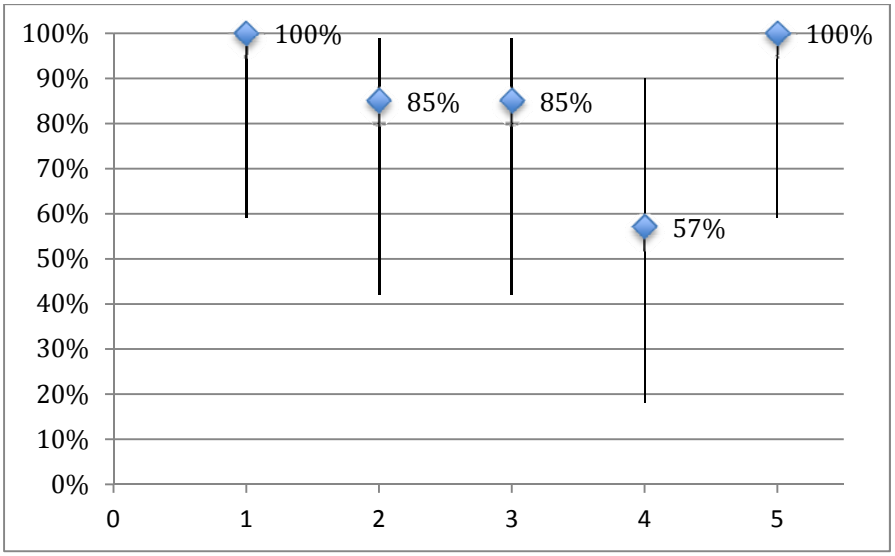

Figure 4: Task 1, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

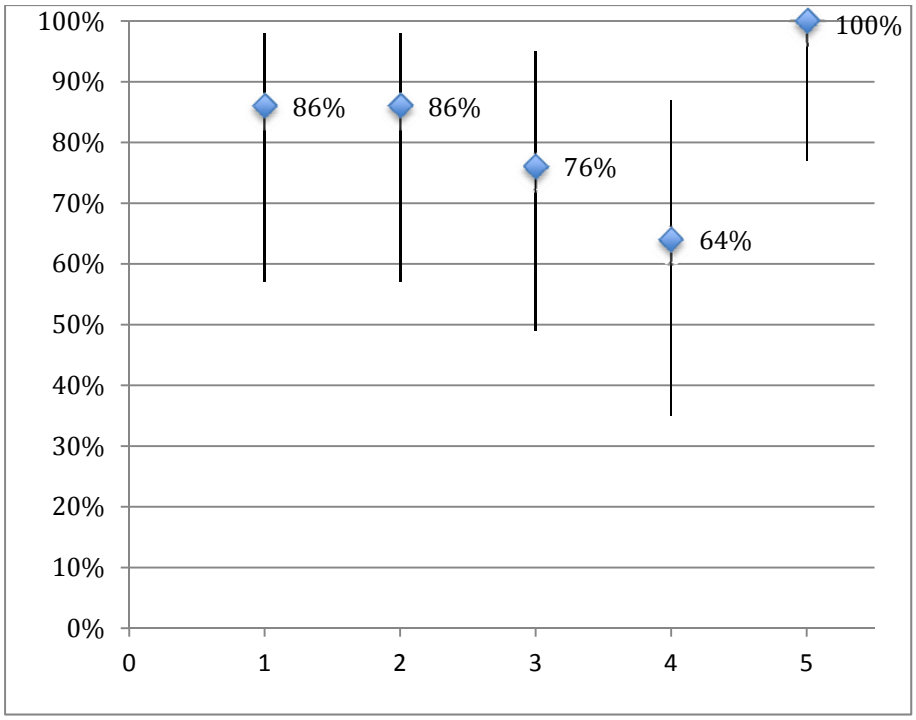

Figure 5: Task 1, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

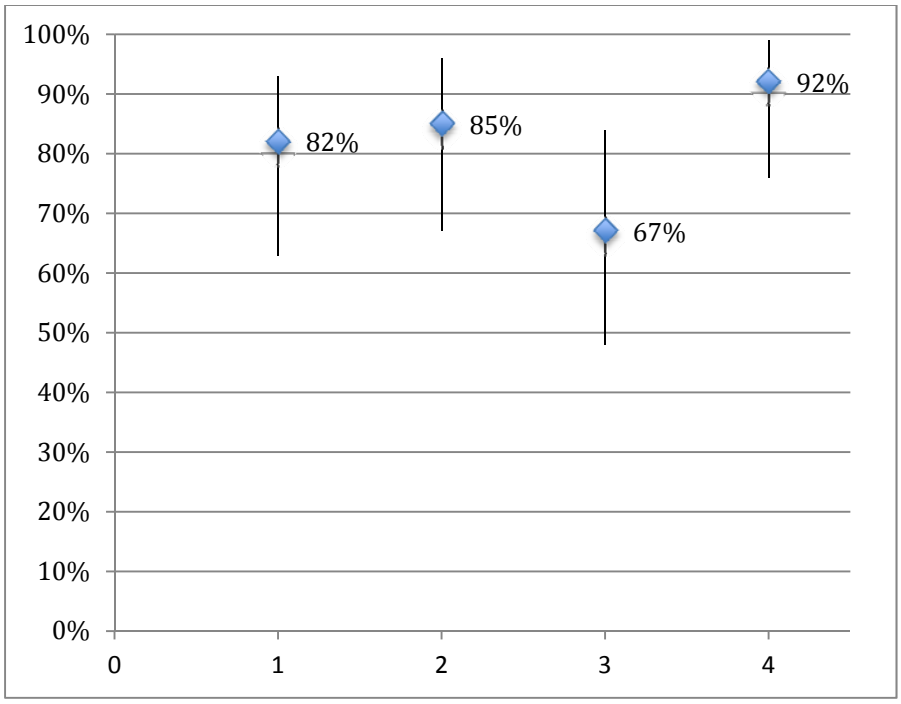

Figure 6: Task 1, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

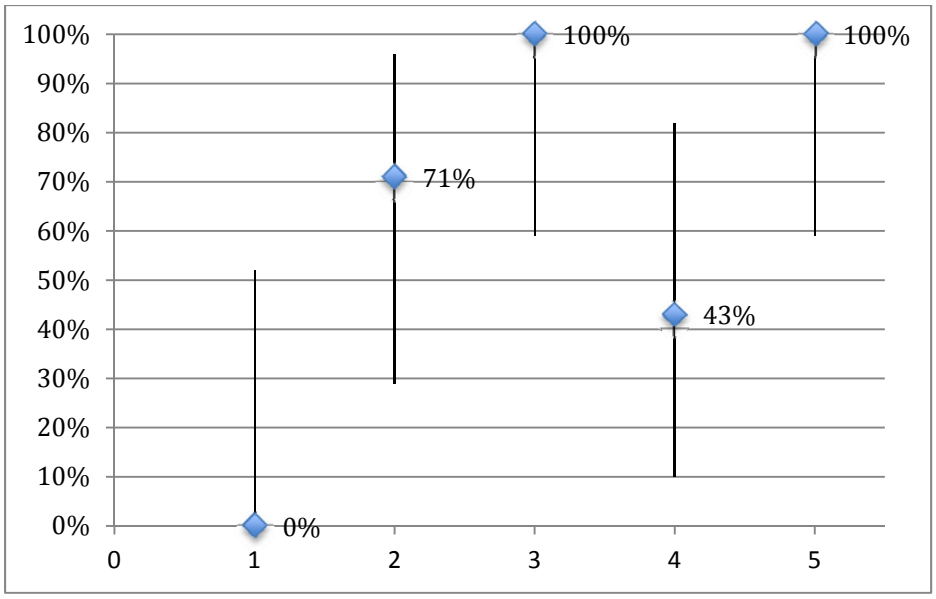

Figure 7: Task 2, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

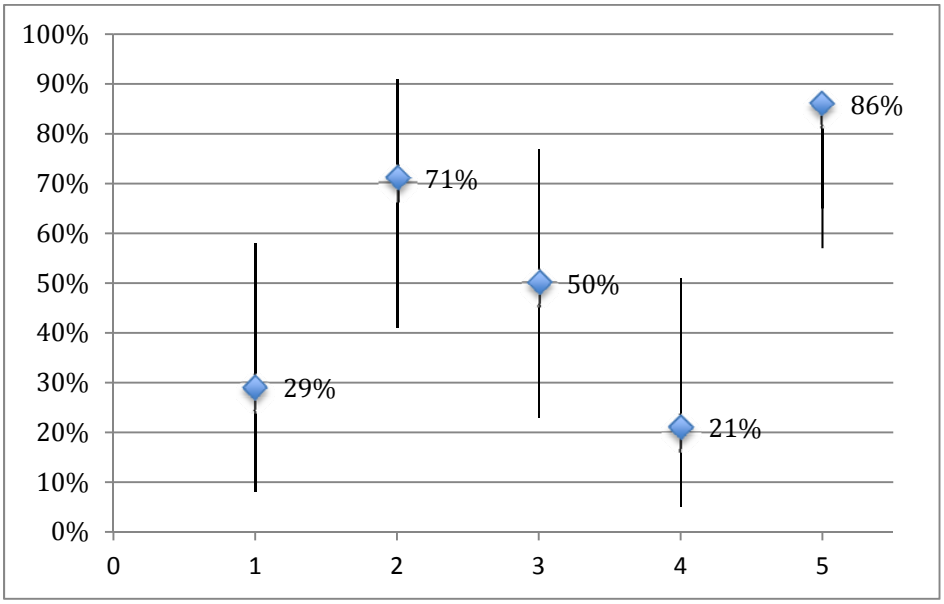

Figure 8: Task 2, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

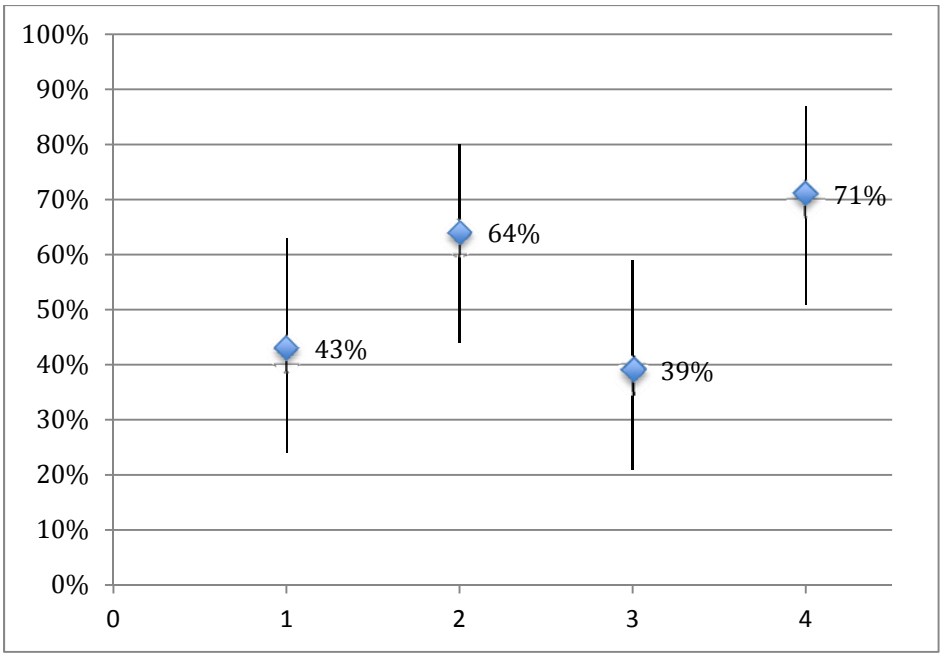

Figure 9: Task 2, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

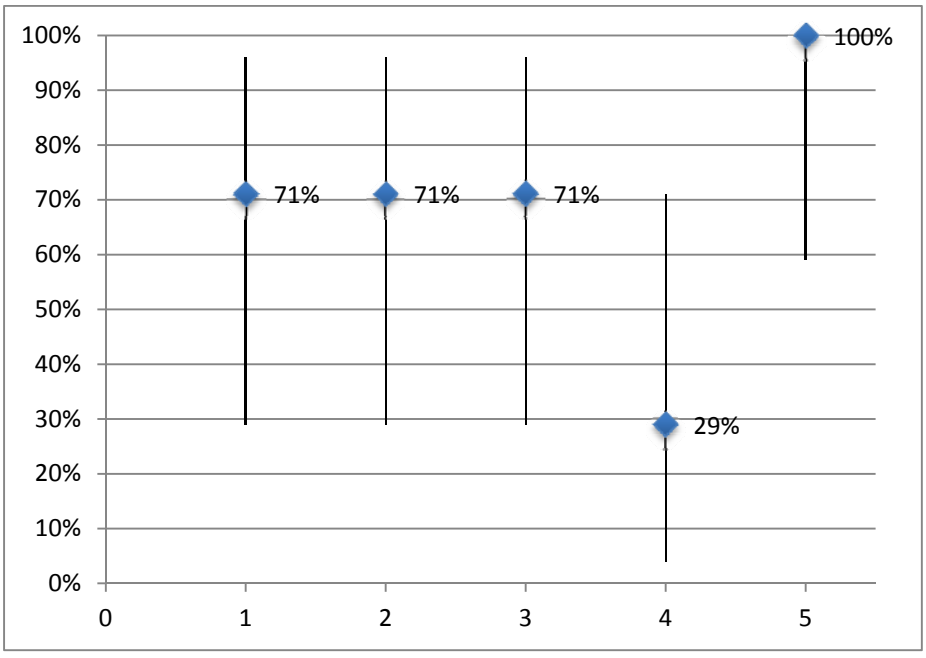

Figure 10: Task 3, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

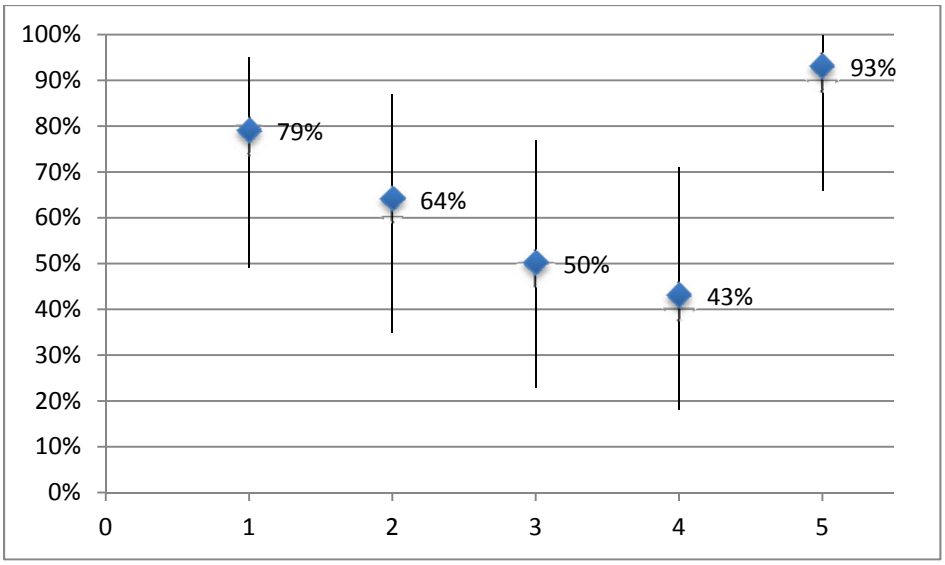

Figure 11: Task 3, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

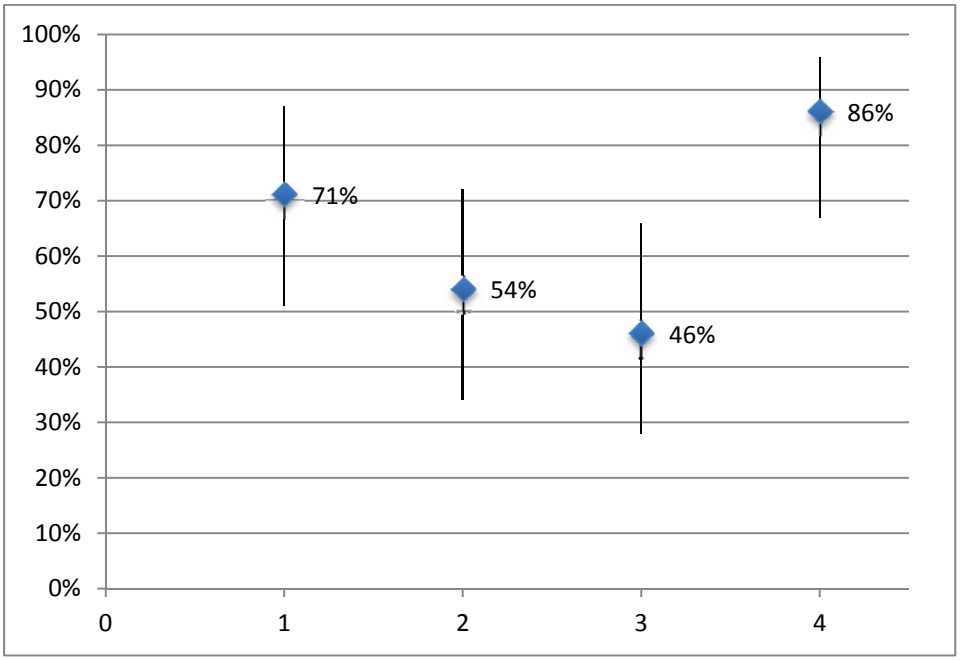

Figure 12: Task 3, Predicted Values for Completion Rate, with 95% Confidence Intervals, for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

It can be seen that though the performance rate varies significantly, with all but one exception (data from Task 2 sampled 7 at a time) the Confidence Intervals for Task Success Rate show significant overlap. Therefore, this test shows reliability at different sample sizes. However, the wide variations in predicted value make the smaller sample sizes problematic for a pass/fail test. The larger sample sizes show more consistency in predicted value and also have a smaller

Confidence Interval. The larger sample size is, therefore, more suitable for a pass/fail criteria-based test. However, pragmatically, larger samples also result in a higher cost to perform the test.

The data from testing was also analyzed using the proposed Time & Success metric following the same resampling procedures described above. The results are shown in the tables below.

Sample	Task 1	Task 2	Task 3
1	63.5%	36.7&	11.4%
2	57.7%	23.3%	4.4%
3	67.3%	48.1%	12.6%
4 (Lowest Possible)	52.2%	9.3%	0.0%
5 (Highest Possible)	87.4%	58.0%	7.5%

Table 4: Calculated Values for the Time & Success Metric for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

Sample	Task 1	Task 2	Task 3
1	67.2%	37.1%	9.0%
2	70.0%	31.9%	9.3%
3	57.8%	31.6%	12.6%
4 (Lowest Possible)	47.2%	10.8%	0.0%
5 (Highest Possible)	72.7%	49.2%	17.9%

Table 5: Calculated Values for the Time & Success Metric for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

Sample	Task 1	Task 2	Task 3
1	68.9%	33.5%	10.1%
2	59.5%	27.9%	8.6%
3 (Lowest Possible)	53.7%	13.1%	5.3%
4 (Highest Possible)	80.4%	44.5%	11.3%

Table 6: Calculated Values for the Time & Success Metric Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

It can be seen from these tables that there is also a wide variation in the Time & Success metric, although smaller in magnitude than the variation in the underlying Task Success Metric it is based on. And, as with the Task Success Metric, this variation can also be seen to be inversely correlated with the sample size (e.g., there is more variability in smaller sample sizes) with variances as high as 49% for samples of 7 participants at a time to 32% or less for samples of 28 participants at a time.

The 95% Confidence Interval was also calculated for each of the resampled groups plus the

lowest and highest possible values. These values are plotted in the figures below.

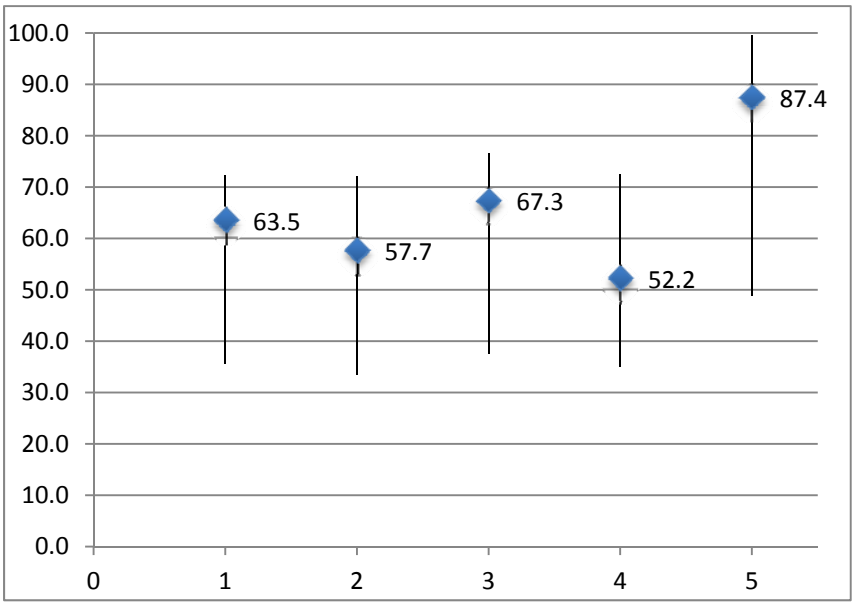

Figure 13: Task 1, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

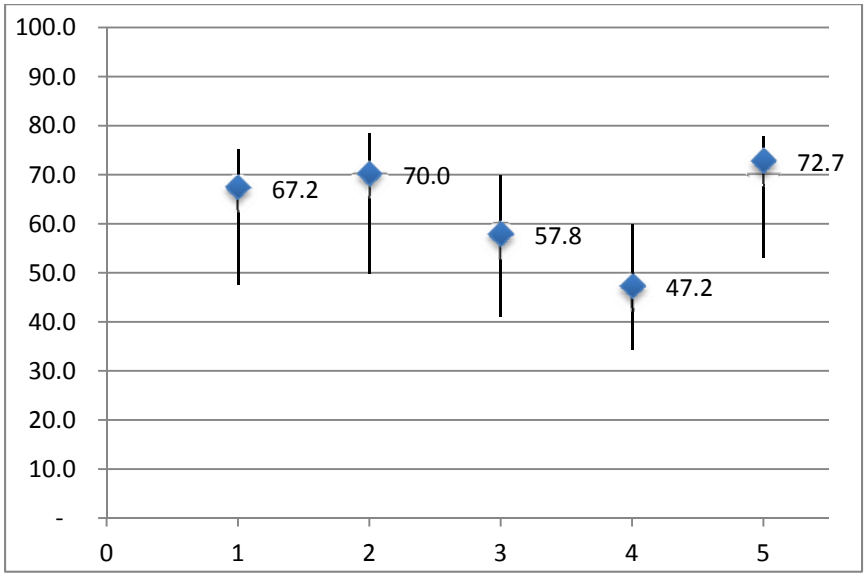

Figure 14: Task 1, Calculated Values for the Time & Success metric,, with 95% Confidence Intervals, for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

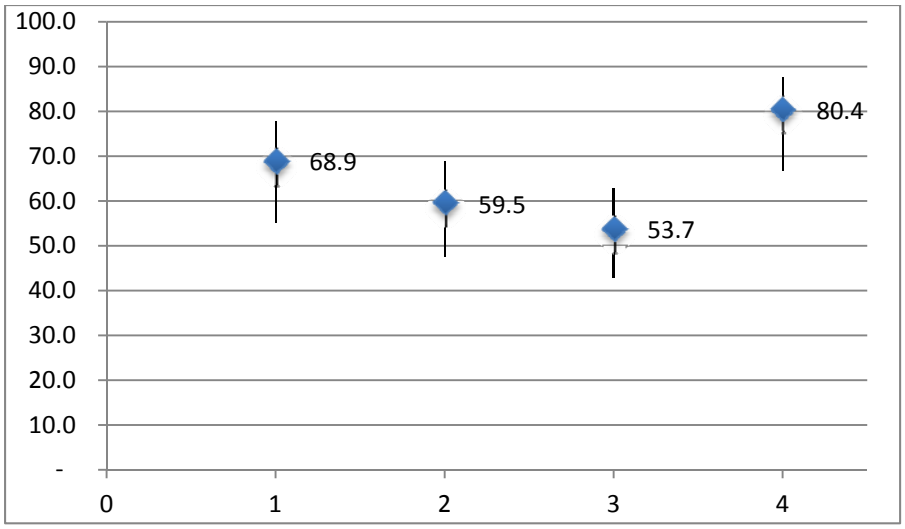

Figure 15: Task 1, Calculated Values for the Time & Success metric,, with 95% Confidence Intervals, for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

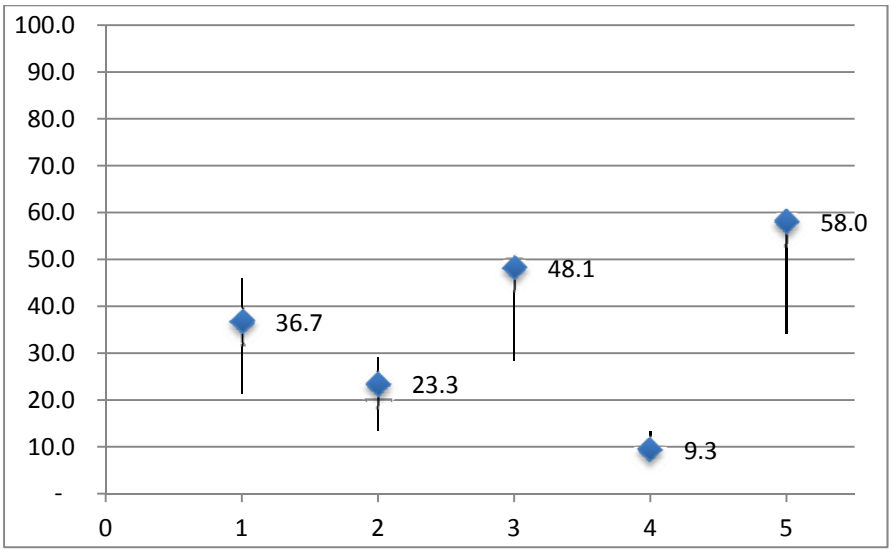

Figure 16: Task 2, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

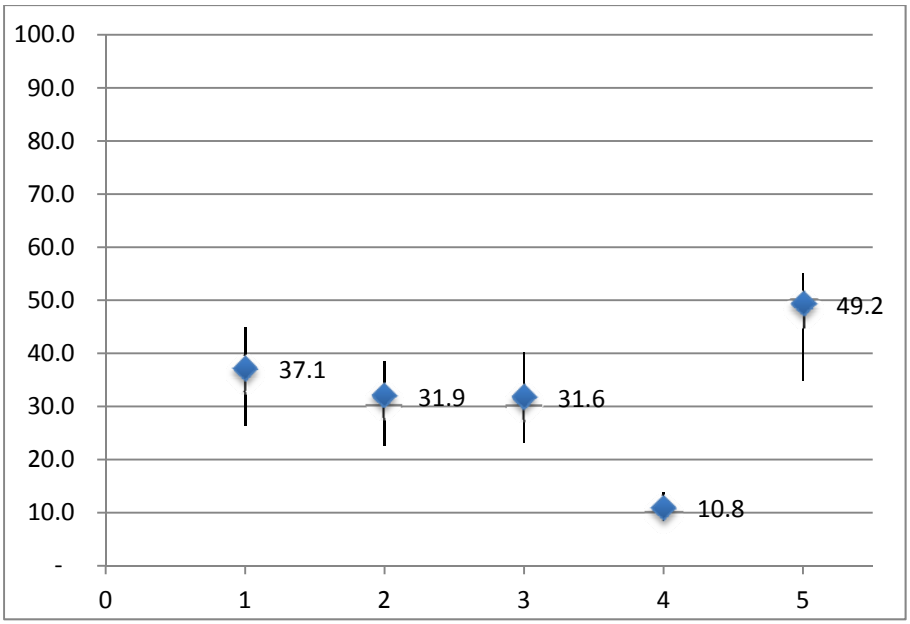

Figure 17: Task 2, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

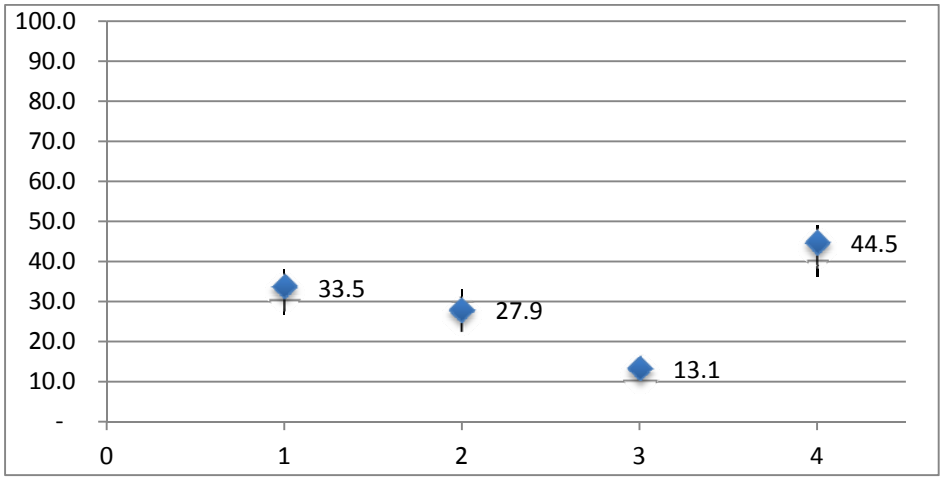

Figure 18: Task 2, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

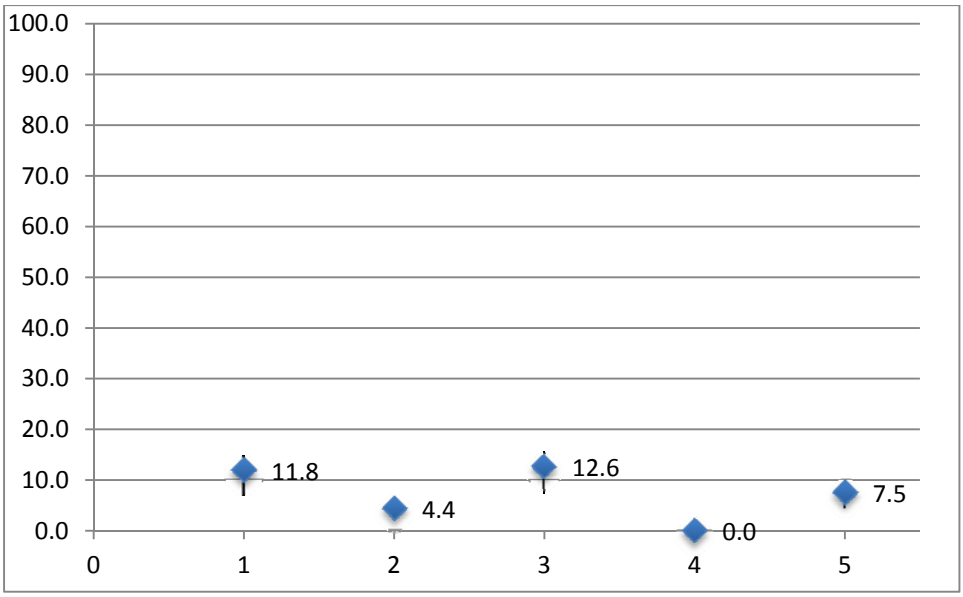

Figure 19: Task 3, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 7 at a Time plus the Lowest and Highest Values Possible from the Sample

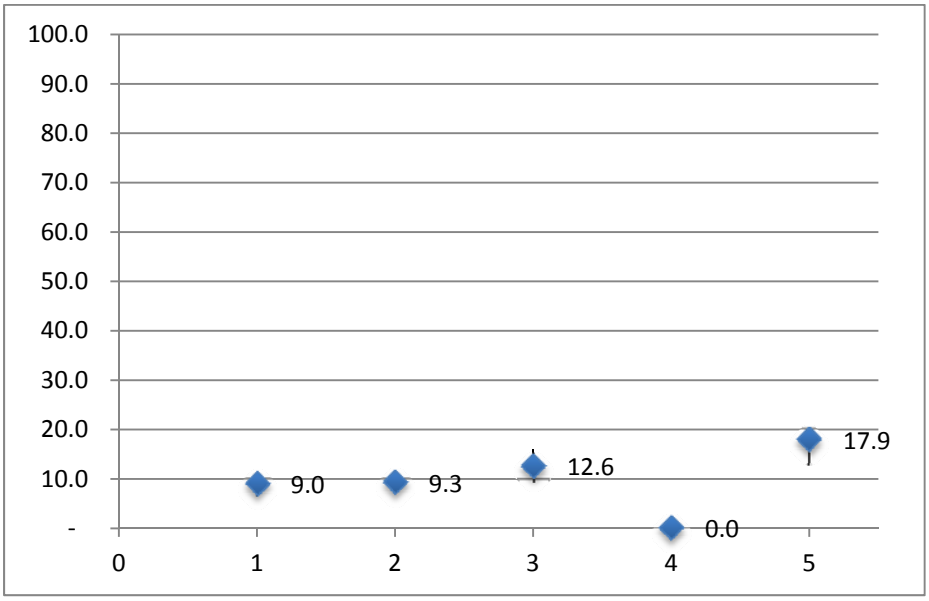

Figure 20: Task 3, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

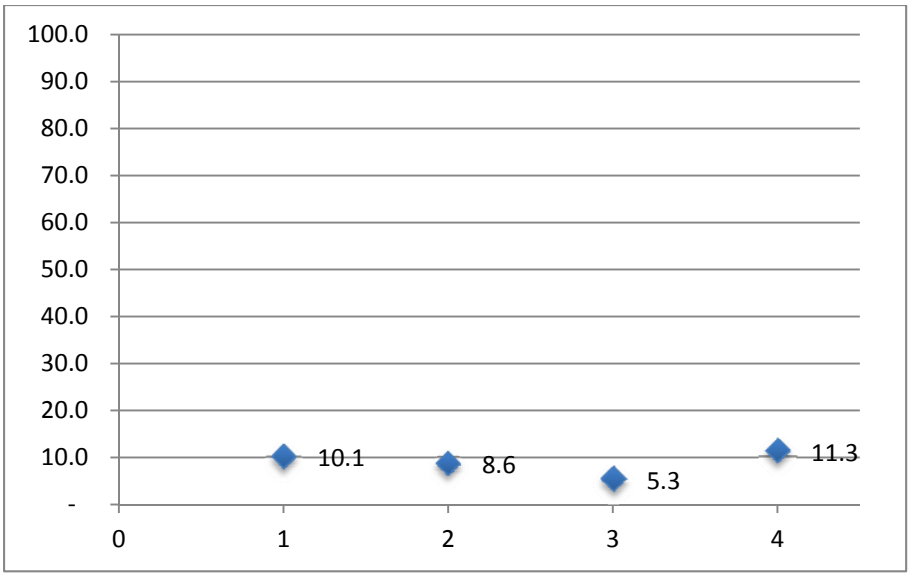

Figure 21: Task 3, Calculated Values for the Time & Success metric, with 95% Confidence Intervals, for Total Population Sampled 28 at a Time plus the Lowest and Highest Values Possible from the Sample

These charts show that there is little or no overlap in the confidence interval around the Time & Success metric, showing that the Time & Success metric has less reliability than the underlying task completion rate upon which it is based. This can be a problem since different results are obtained on the test even for a sample size of 28. For example, for Task 2, the highest Time & Success metric obtained from the sample population was 44.5%–a passing score based on the spreadsheet criteria provided in the RFP. However, the three other values obtained for the Time & Success metric for Task 2 were all below the cut off value for this task even when the confidence interval is included.

Box plots for Task 1 are provided below to help understand the variability in the proposed Time & Success metric.

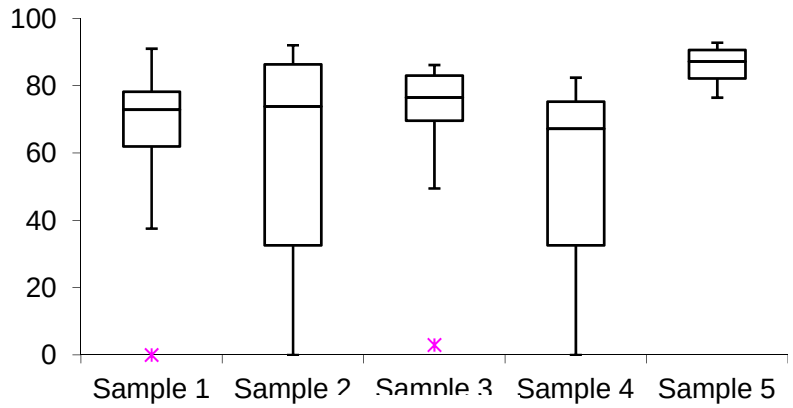

Figure 22: Box Plots, Task 1, Participants Sampled 7 at a Time

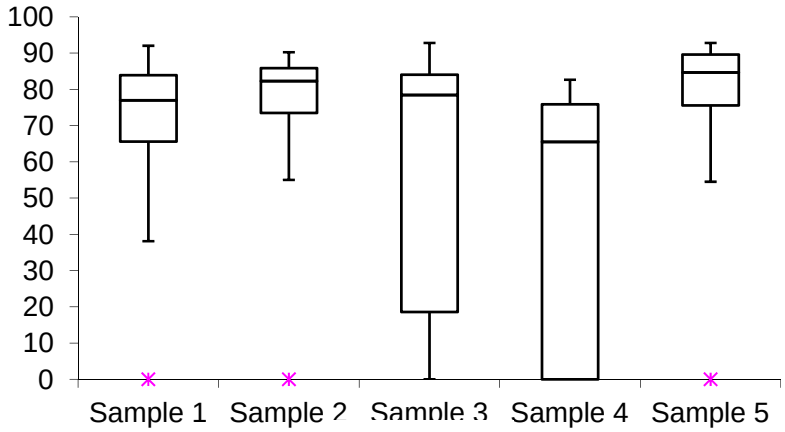

Figure 23: Box Plots, Task 1, Participants Sampled 14 at a Time

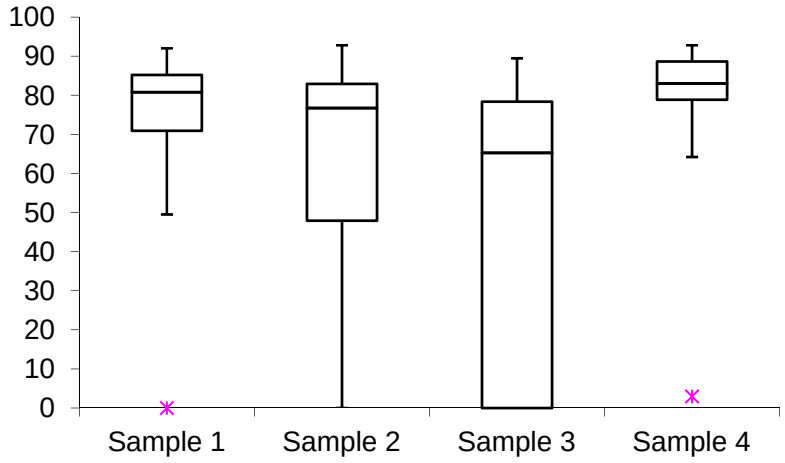

Figure 24: Box Plots, Task 1, Participants Sampled 28 at a Time

The proposed metric is also based on calculating the arithmetic mean of the Time & Success scores for each participant. Histograms of the Time & Success metrics are also provided below.

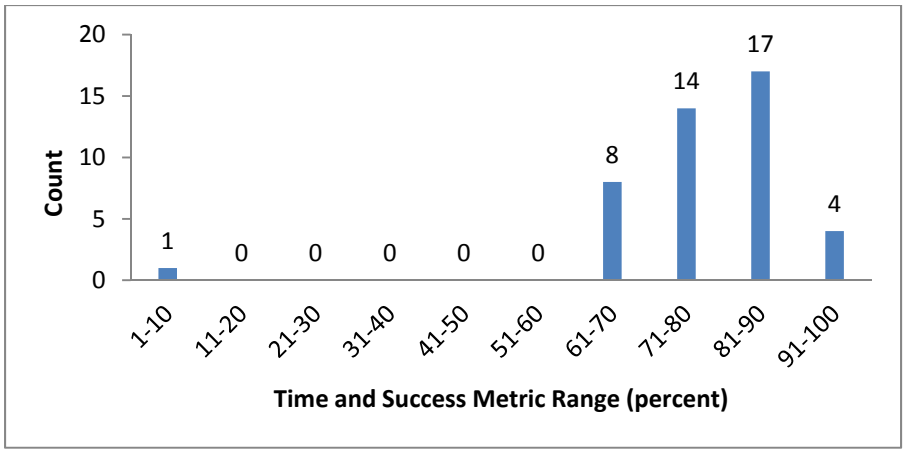

Figure 25: A Histogram of the Time & Success Metric, All Participants, Task 1 (Non Zero Values Only)

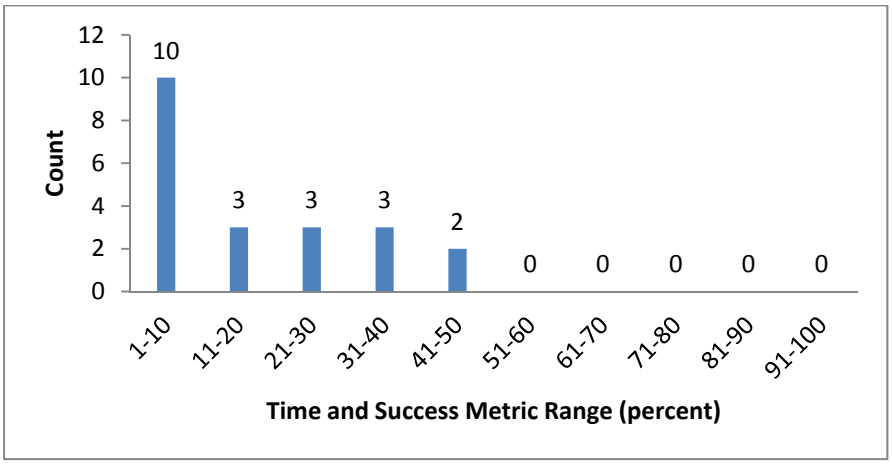

Figure 26: A Histogram of the Time & Success Metric, All Participants, Task 2 (Non Zero Values Only)

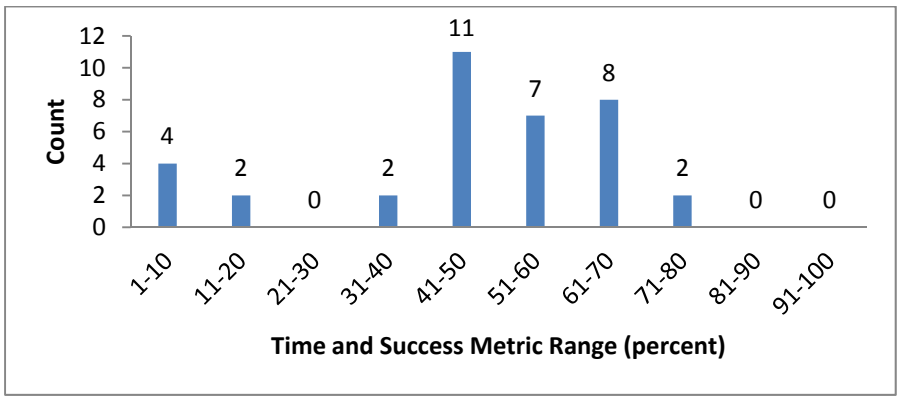

Figure 27: A Histogram of the Time & Success Metric, All Participants, Task 3 (Non Zero Values Only)

The Time & Success metric data for each task, also analyzed using the normal quantum tile plot method, was determined to be non-normally distributed. Therefore, a calculation of the arithmetic mean of this data is not considered to be statistically valid. If the Time & Success

metric is to be used the geometric mean should be considered.

3) Suggestions for Minimizing Variability in:

a) How the Test is Performed

In reviewing the proposed test protocol prior to conducting this research, two procedural elements were identified that pose a threat to the test's internal validity. While performing the dry run of the test, it was noted that participants might glance at the unit under test when reading the test description. For the task of reading the current temperature and set point, this could be a problem. Therefore, an additional instruction was added to the test protocol to have the participant's face away from the unit under test when reading the task descriptions.

Secondly, task timing was dependent on the participants announcing that they were done with that task. While performing the dry run of the test, it was noted that participants might forget to announce they are finished with the task, particularly when the task is very short. Therefore, an additional instruction was added to the test protocol to begin and end task timing based on direct observation of the participant. Since the earlier change was to have the participant face away from the unit while reading the task, the task start time would be when the participant turns to face the unit. The task end time is slightly more subjective and could be based on the participant announcing they are done with the task or when they turn around to read the next task.[1]

Finally, it is recommended that all test material be presented in written format and reviewed prior to all device interaction. Once testing has started, the test administrator should leave the test room to minimize observer effects and to limit interaction between the test participant and the test facilitator. The test administrator should be provided with a specific script to use if questions are asked (e.g., "I'm sorry, I cannot help you once the test has started. If you are stuck on a task, you can move on to the next one.")

b) The Test Results

Assuming procedural issues are addressed to ensure consistency in how the test is performed, variability of the test result will be based on the sample size and sample variability. See the discussion on recommended user group composition and size for a discussion on how to minimize test data variability based on these factors.

4) Suggestions for Improving the Test Script

In reviewing the initial test script provided at the outset of this research, recommendations

[1] To avoid falsely assuming task completion, the perceived end of the task is noted and then confirmed with the participant. This allows for cases were participant might appear to be done or might turn to ask a question of the test facilitator.

were developed that were considered necessary to improve the test script. Most of these problems and appropriate remediation were presented to the committee earlier in this project. These recommendations addressed three areas of the current test script.

a) Use of Unfamiliar Terms

The test description should avoid the use of uncommon terms. For example, the originally proposed test protocol used the term "active setpoint" to describe the temperature setting in use. This term may be unfamiliar to some participants and may cause confusion. It is recommended that more universal terms be used. For example, the originally proposed wording for Task 2 was:

> *"…please read aloud the current room temperature and the set temperature, also called the active setpoint*"."

This task was reworded to read:

> *"…please read aloud the current room temperature and the temperature the Residential Climate Control unit is trying to maintain."*

Inclusion of Non Task Specific and Background Information

Test descriptions should be short and precise including only information necessary for the task. Background and explanatory information included in a test description can distract users from the task. For example, the originally proposed wording for task 5 was:

> *"The Climate Control is controlling heating in Program Mode. In this Winter mode, room temperature is controlled according to a schedule to maintain comfort when the home is occupied and to save energy when occupants are away or sleeping. Climate Control program schedules may be adjusted to meet your personal or family's schedule. You would like your home to be automatically heated to a comfortable temperature all day on Saturdays. For this task you will be asked to adjust the Climate Control to make this change…."*

This task was reworded to read:

> *"The Residential Climate Control unit is currently controlling heating. It is using a predefined schedule for heating. You want to adjust the Residential Climate Control to include the following changes…"*

b) Order of Tasks

The original proposed test protocol stated the tasks were ordered from easiest to hardest. Though there is rationale for this approach, other compelling factors should be taken into consideration. For example, the tasks should have a logical order to a new user. This is likely a

more compelling criterion for setting the task order. The more difficult task of initial set up could be addressed prior to a daily operation task. This provides a logical order reflecting how the device will likely be used. It also provides an opportunity to develop a conceptual model of the device that users need for daily operations tasks. If concerns remain, participants could be told that the tasks are separated into the initial tasks and the others. This is likely sufficient information for the participants since it meets logical expectations that initial set up tasks on a device are more complicated than daily operations. However, there was no data from testing that suggested this additional separation is needed. It is likely the participants understood which tasks were initial set up tasks from their titles. The most relevant consideration is to ensure that tasks referring to other tasks are performed in the correct order. For example, the test protocol includes programming automatic settings as well as overriding these automatic settings. In the original proposed protocol, participants were asked to override the program settings before they had programmed them (since programming was more difficult). However, since participants lacked a conceptual model for the override task (unless they had experience with programmable thermostats), the override task was ambiguous; it could be accomplished either by a temporary override or by reprogramming the unit. In this research, the order of these tasks was reversed and the second task (simply setting the unit to on and to heat) should include a reference to the programming task (an instruction to complete the task without reprogramming the unit).

c) Use of Compound Tasks

Task 2 required people to read the current temperature and to read the temperature to which the unit was set to maintain. This task appeared to be problematic for some participants based on the unit design (participants reversed the meaning of the two display elements). However, other participants appeared to have difficulty understanding the task itself. Some participants modified the set point to match the current temperature, demonstrating they likely understand the two display elements. However, since subjective interpretation should not be included in a performance-based test, their performance would be considered task failure based on the task wording. It is recommended that each task be related to a single goal (e.g., reading the current temperature).

5) Recommendations for Appropriate User Group Composition

The initially proposed test protocol included variations in demographics so that the user group was “representative of U.S. demographics on a National basis.” Though this demographic does provide greater face validity for the test, variations in demographics threaten the validity of the test. Also, since the demographic factors are specified as independent criteria for recruitment, combinatorial factors can also threaten the test’s reliability.

Since the test is a comparison between the results for a unit under test against pass/fail values,

internal validity of the test (the validity of the test itself) can be maintained using any population for testing provided that the pass/fail criteria are based on the same population. However, the results cannot be generalized beyond the characteristics of the population used. A more strictly defined population would increase the test reliability and would potentially allow for fewer participants to reach the same confidence level. However, this more narrowly defined population would reduce the external validity of the test (the ability to generalize the results). A population could still be selected to increase the face validity of the test, such as the population most likely to have problems or the population most likely to use the product. Alternately, one could simply choose the population that is easiest to recruit. Additional analysis could be performed, if it is decided it is necessary, to determine how the selected population equates to a more representative population, but this exercise is not needed to address test issues of reliability or construct validity. However, it should also be noted that the actual population of users of these devices is not currently known, so generation beyond the test population may not even be possible unless this data can be obtained.

An example of the effect of variation in population characteristics on the results is shown below. Using the current population demographics sampled 14 at a time, there was a 36% variance in the point estimate (and a 13% variance in the values obtained for the upper confidence interval) for Task 1 – setting the current time in the Task Completion Rate (see the table below).

Sample	Task 1	Upper CI Value
1	86%	98%
2	86%	98%
3	76%	95%
Lowest Possible	64%	87%
Highest Possible	100%	100%

Table 7: Predicted Values for Completion Rate and Upper 95% Confidence Interval Value for the Original Population Sampled 14 at a Time plus the Lowest and Highest Values Possible from the Sample

This population includes 3 levels of education – less than HS, HS to Less than BS, and BS or Higher. When the population of participants was resampled but restricted to an education level of HS to less then BS (maintaining an equal distribution of ages and gender), there was only a 17% variance in the point estimate (and a 2% variance in the upper confidence interval), even when sampled at a population of only 12 at a time (see the table below).

Sample	Task 1	Upper CI Value
1	91%	99%
2	100%	100%
3	100%	100%
Lowest Possible	83%	98%
Highest Possible	100%	100%

Table 8: Predicted Values for Completion Rate and Upper 95% Confidence Interval Value for the A More Homogeneous Population Sampled 12 at a Time plus the Lowest and Highest Values Possible from the Sample

6) Recommendations for the Appropriate Group Size

Overall, Task 1 showed the highest performance data with 43 of the 50 participants completing the task successfully. Task 2 showed the next highest level of performance with 40 out of 50 participants completing the task successfully. Task 3 showed an overall performance level of 30 out of 50 participants completing the task successfully. As previously mentioned, Task 4 showed such low performance data that is not included in any analysis.

7) The Use of Confidence Intervals for Determining Pass or Fail

Once a test protocol is developed that shows statistical reliability, it is also recommended that the Confidence Interval (CI) be included with the point value for all pass/fail criteria. Since all testing is an estimate based on a sample, the value obtained is unlikely to ever exactly match a previously obtained value. The Confidence Interval is the only way to show reliability of the test. It is further recommended that the 95% Confidence Interval be used. To provide the "benefit of the doubt" to the vendor, the upper Confidence Interval level should be used to determine if the unit passed the pass-fail criteria for any task on the test. A unit under test would pass the test if the upper value of the CI is equal to or greater than the pass/fail cutoff point for the task on the test.

It should be noted that the CI also lessens the ability to detect small differences between units of different designs, if that is a concern. The CI can be reduced by lowering the confidence level required (e.g., using a 90% confidence level rather than a 95% confidence level) or by increasing the sample size. A larger Confidence Interval could allow a design to pass on performance data (i.e., for a cutoff point of 94, a value of 85 with a CI of 18, i.e., $\pm$ 9, or greater would pass). Therefore, the question of the appropriate group size is partially dependent on what is considered to be an acceptable CI both to vendors (mostly concerned with the confidence level) and to the government (mostly concerned with the CI size).

The general rule of thumb for research involving human subjects is that a population of 25-30 is needed to ensure the data shows regression to the mean.[2] **It is, therefore, recommended that a**

[2] For example, Cohen suggests 30 participants should lead to about 80% power (the minimum suggested power for an ordinary study) given a medium to large effect size. See Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

population size of *at least* 25 be used. It should be noted, however, that this also assumes a homogenous population, so it is recommended that the demographic definition be redefined to minimize performance variations as discussed above. The recommendation for a sample size of *no less than* 25 is based solely on industry standards as the *minimum* number of participants. At a 95% confidence level, this would provide a Confidence Interval of approximately 20-30% (depending on the point estimate value). Since only half of the Confidence Interval is above the point estimate, this provides a 10% to 15% margin of error on the test. However, this value may not be sufficient to discriminate between devices with similar performance levels. If this is needed, additional participants may be needed to further reduce the CI. The cost of the test is, in large part, directly proportional to the population size, so a balance is needed between total cost and desired discrimination level of the test.

8) Recommendations for Maximum Time to Complete Tasks

Is maximum time to complete tasks necessary?

The inclusion of a maximum time-on-task would be warranted provided time-on-task had a direct correlation with the ability to successfully complete the task. This is generally not the case. For example, in the data from this research for Task 1 (the only test with normally distributed data) is shown in the table below.

Sample	Performance	Mean TOT
1	100%	90
2	85%	34
3	85%	56
Lowest Possible	57%	93
Highest Possible	100%	72

Table 9: Task 1, Completion Rate and Time-on-task (TOT) Data for Total Population Sampled 7 at a Time with the Lowest Possible and a Highest Possible values

Sample	Performance	Mean TOT (sec)
1	86%	81
2	86%	53
3	76%	86
Lowest	64%	115

Possible		
Highest Possible	100%	80

Table 10: Task 1, Completion Rate and Time-on-task (TOT) Data for Total Population Sampled 14 at a Time with the Lowest Possible and a Highest Possible values

Sample	Performance	Mean TOT
1	82%	85
2	83%	83
Lowest Possible	67%	101
Highest Possible	92%	73

Table 11: Task 1, Completion Rate and Time-on-task (TOT) Data for Total Population Sampled 28 at a Time with the Lowest Possible and a Highest Possible values

The tables above show there is no correlation between the mean time-on-task and the task performance rate. Therefore, the maximum time to complete tasks is not recommended as an approach to test a unit's usability.

Though a maximum time-on-task for each task is not recommended, an overall maximum time to complete all tasks in the test is recommended to ensure that testing remains on schedule. Since an externally defined limit may not correlate with task completion, this maximum time is not directly associated with product usability but is associated with the practicality of the ensuring that the entirety of the test remains on a given schedule.

9) If so, What should the Maximum Time to Complete each Task be Documented in the Test Procedure as?

Maximum time per test is not recommended. (See the discussion above.)

10) Estimated Cost for Product Qualification Testing

The cost for a test of this nature includes both recurring and nonrecurring costs. The recurring costs are directly associated with the number of test participants and include the recruiting cost, stipend cost, and cost for data collection. Since the test procedures and reporting procedures are predefined, and since the data analysis can (and should) be programmed, the recurring costs are minimal. Using normal values for recruiting, stipend, and labor rates the estimated cost for a test with a population of 25 participants is estimated to be approximately $10,500. However, if a reference unit is required (see discussion below), the population doubles to 50 participants but

the test cost only increases to approximately $18,000.[3] An additional day is likely required to allow for the testing staff to become familiar with the product and adapting the test procedures to the specific unit under test. Therefore, the total cost may increase by approximately $2,000.

11) Expert Opinion on the Value of Parallel Testing for a Reference Device and the UUT to Improve Test Repeatability, with Recommended Test Method Implementation Strategies for a Reference Device

The use of a reference device would not have an effect on test repeatability, but its use would affirm the validity of a specific test run. Since failure to adequately follow all procedures correctly could invalidate the results of a specific test, the use of a reference unit would demonstrate that all procedures (recruitment through analysis) were correctly followed if the test returned the expected value for the reference unit.

From a research standpoint, the reference unit could be any unit whose value on the test is already known (i.e., any unit that has previously been tested using the protocol). However, to avoid any appearance vendor bias, a simulated unit with equivalent functionality could be used in lieu of an actual device. The reference unit would also serve to test for "test drift" based on changes in the population over time.

Using a reference unit approach, a total population would need to be recruited that is twice the required number for testing a single unit. The participants would then be randomly assigned to either the UUT or the reference device for the testing, while maintaining the correct demographic distribution in each group. Both units would then be tested using exactly the same procedures and at the same time. Since the value of the second unit is known prior to testing, the results of testing should result in this expected value (within the Confidence Interval of the test) as a way to show that the recruitment, operational procedures, and analysis have been correctly followed. This provides additional assurance that the test results for the UUT are valid; however, the use of a reference unit would almost double the cost of testing.

The need for this additional assurance that the test results are valid is dependent on an agreement between the government and vendors. If vendors are willing to abide by the findings of the test given the statistical uncertainty inherent in this type of testing (even at 95% confidence level), then a second, reference unit is not needed. However, if vendors need additional assurance that the results of the test are valid, the use of this reference unit is recommended.

[3] These estimates assume two staff members and a rental cost for a test facility.

12) Survey Data

A survey was conducted after testing. Though other questions were asked about adjustments, they were asked about their experience with programmable thermostats (as well as other products) and their self-reported knowledge of HVAC systems. And analysis was performed of survey responses against the performance data to determine if there was an interaction effect.[4]

The typical approach of this type of analysis would be a cross tabulation test using a chi-squared analysis. However, the sample size in this case was too small to yield statistically significant values. Therefore, the data was graphed to look for trends in the data. The graphs for performance time based on level of experience with programmable thermostats are shown in the three tables below.

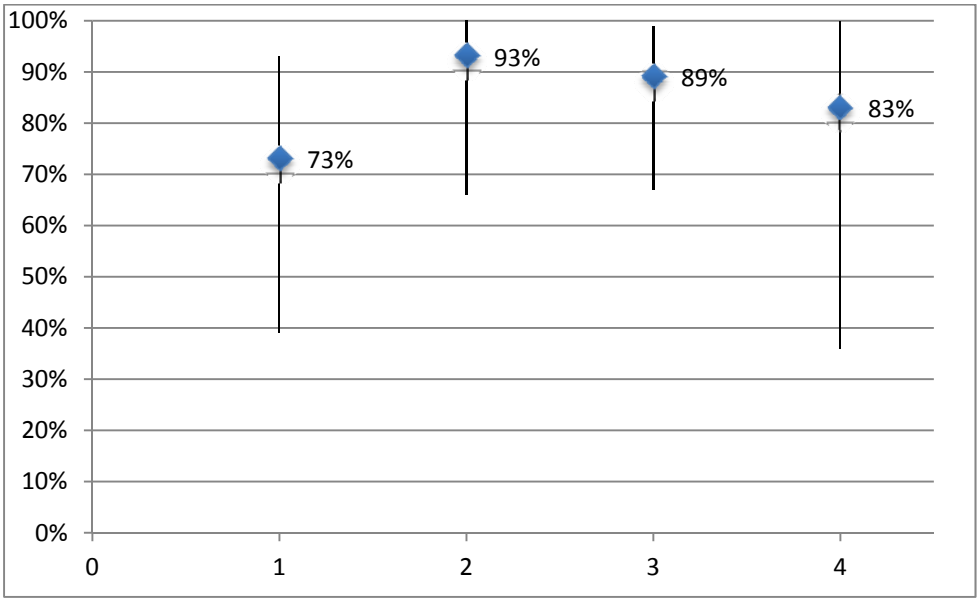

Figure 28: Task 1 Completion Rate based on Prior Experience with Programmable Thermostats (1=Low Experience)

[4] This analysis was restricted to pass/fail performance rate data since the Time-on-Task data was not normally distributed and did not correlated with the performance data.

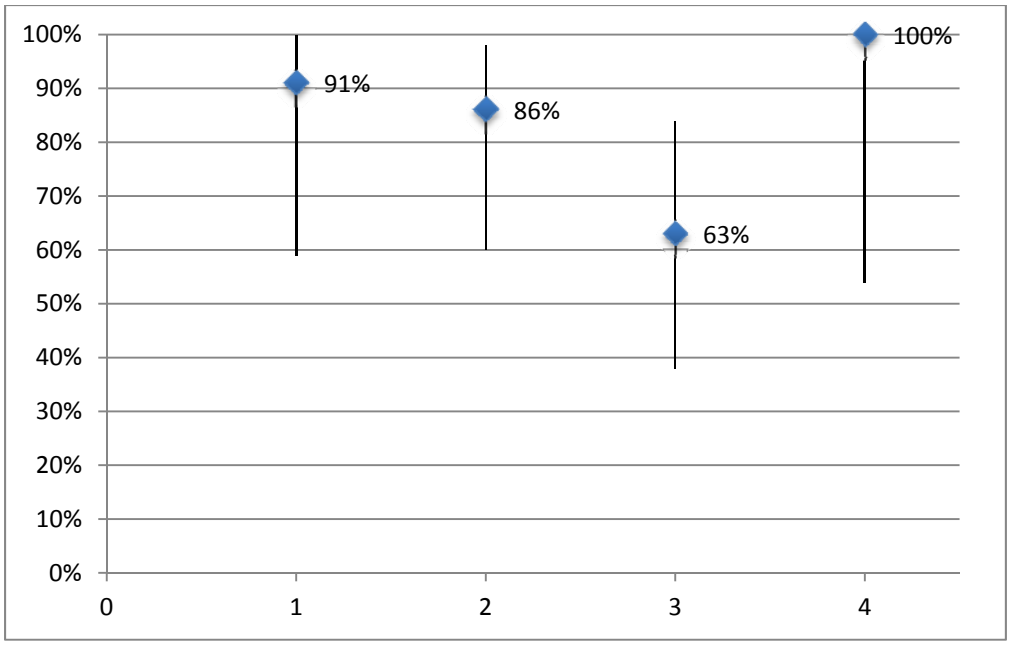

Figure 29: Task 2 Completion Rate based on Prior Experience with Programmable Thermostats (1=Low Experience)

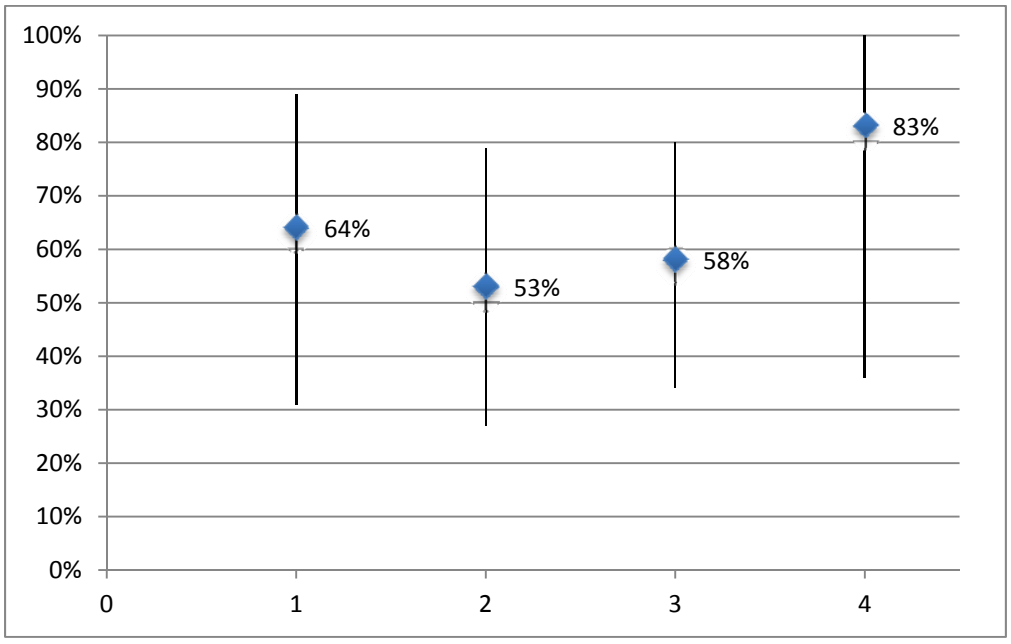

Figure 30: Task 3 Completion Rate based on Prior Experience with Programmable Thermostats (1=Low Experience)

As can be seen from the table, the data suggest that there is no correlation between the participants' self-reported experience with programmable thermostats and their performance on any of the first three tasks.

The data was also analyzed based on the person self-reported knowledge of HVAC systems. The grass for these analyses can be seen in the three graphs below.

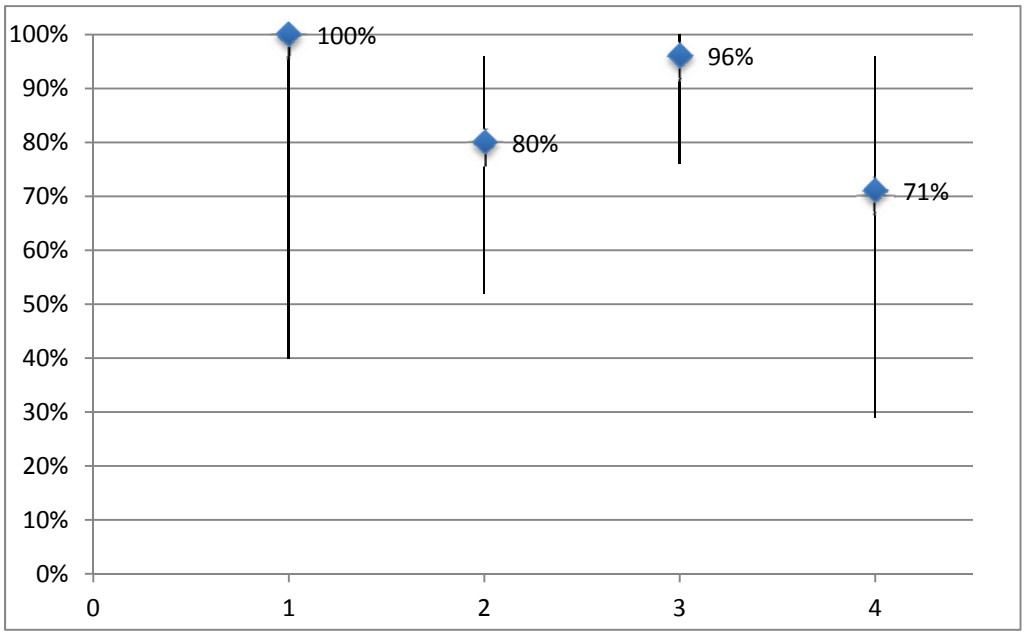

Figure 31: Task 1 Completion Rate based on Self Report knowledge of HVAC Systems (1=Low Knowledge)

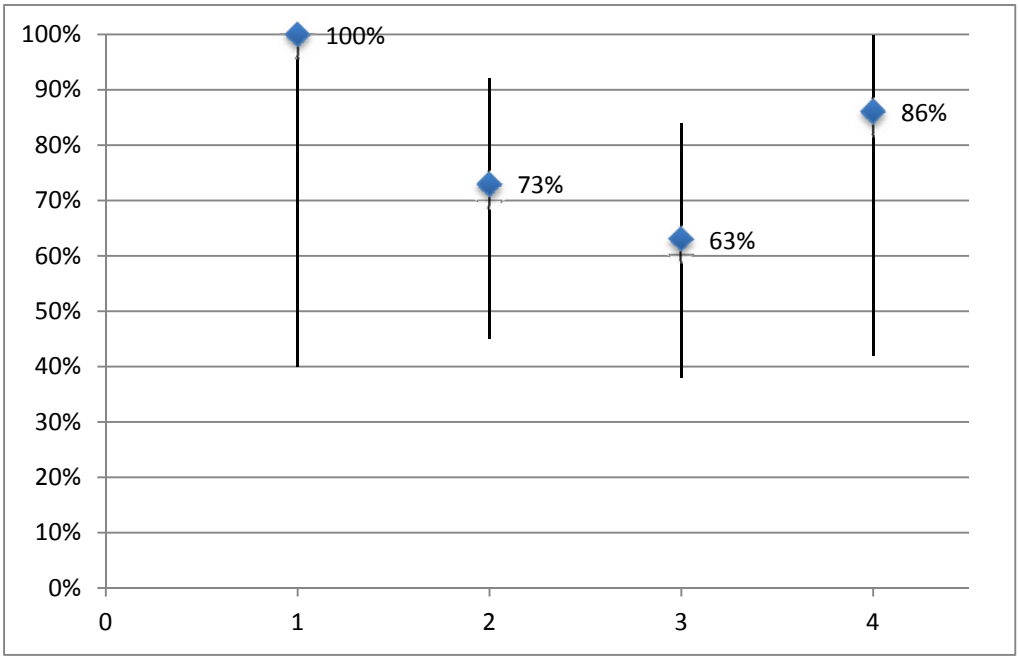

Figure 32: Task 2 Completion Rate based on Self Report knowledge of HVAC Systems (1=Low Knowledge)

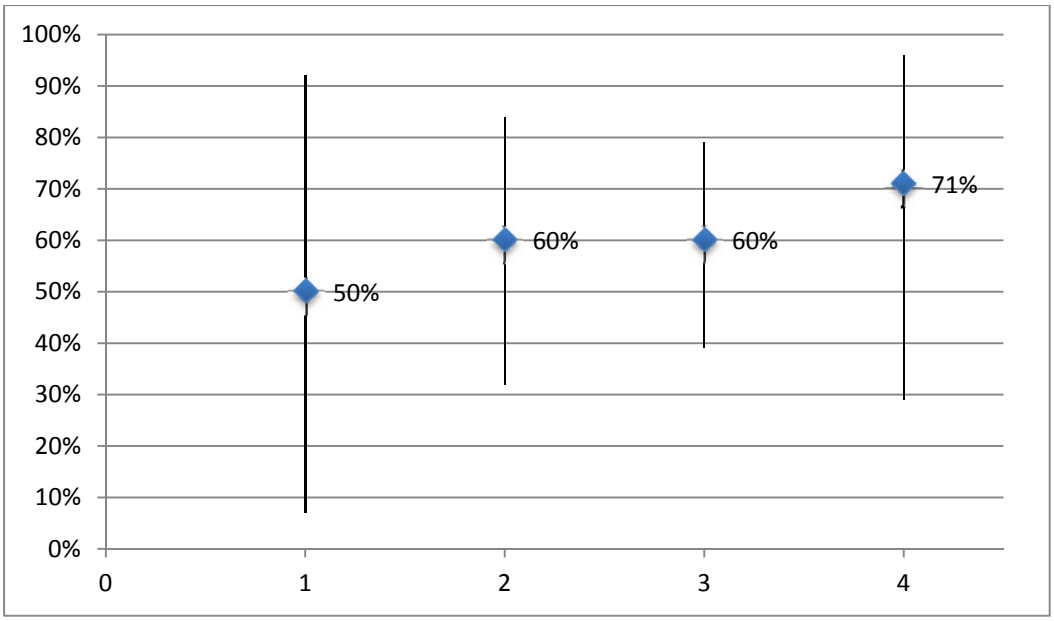

Figure 33: Task 3 Completion Rate based on Self Report knowledge of HVAC Systems (1=Low Knowledge)

As can be seen from the graphs, only task 3 appears to show any correlation between task performance and self-reported knowledge of HVAC systems, but the small sample size precludes drawing this conclusion with any certainty.

13) Other Relevant Information and Professional Opinions to Improve the Test

a) Testing Color Blind Users

The proposed test included a single, red–green colorblind user. The use of a single individual representing a user group is not considered valid. It is **recommended that the test for a colorblind safe design be performed by inspection** and not by a performance based test.

b) Universal Tasks

There are also some considerations that need to be made in the development of the tasks to ensure the tasks are universal for all expected designs; this requires an awareness of the functional abilities for units of this type. For example, a task included in this research asks the user to set the programmable schedule for the unit at a time value that was on the half-hour. This allows for testing of devices that allow programming segments on the ½ hour, 15min., or 10 min. intervals. However, unless it is specifically stated in design requirements set forth by an authoritative body, there is a possibility that a vendor could design a unit restricting users to the setting programmable intervals to the hour.

Similarly, when developing the requirements for setting the time, it is important to consider that some devices may have independent settings for the hour, minute, AM/PM, and date (or day of the week). Other units may combine them together allowing for adjustments of hour and minute, a separate adjustment for AM/PM, and the third adjustment for the date. Other combinations are also possible. For the UUT in this research, the date is independent from the time of day, but the AM/PM designation is linked directly to the clock. In addition, the hours and minutes are linked to each other. The clock is changed by a single control. In the task used in this research, the clock was set back 7 hours and 20 min. This resulted in all participants performing essentially the same task regardless of the time of day when they were tested. However, if the AM/PM designation had been an independent setting, the task would need to have been set to be a minimum of 12 hours behind to ensure that all participants in a full day of testing would be required to set both the time of day and the AM/PM setting correctly. Otherwise, some participants on different units would be required to perform fewer steps for the same task. This would adversely affect the validity of the test when two units were compared. The universal task would be to set the date and time in such a manner that all participants would need to adjust the date or day of week, hour, minute, and AM/PM designation.[5]

c) Common Criteria for All Units

It is also important, if the test is intended to create a pass/fail criteria across units, the units must be compared on equivalent functionality and equivalent tasks. The proposed test procedures include the option to skip a task if a device doesn't support it. It is **recommended that a mandatory set of functions for all units be defined for a pass/fail test**. If a UUT includes a function but eliminates user involvement (e.g., a unit that reads the current date and time from the atomic clock), the unit should pass this element of the test.

Also, since the pass/fail criteria are based on the CI, and the CI is dependent on the sample size, it is **recommended that all tests be based on identical sample sizes**. This may require over recruiting to account for dropouts, or the removal of data deemed invalid for some reason.

In addition, the initially proposed test protocol included providing the participants with the user's manual for reference during testing. Participants were then instructed that they could refer to the manual if desired. This introduces a variable into the test that could affect the reliability of the test, particularly if mean time-on-task data is included. It is **recommended that a consistent procedure be used to address the use of the manual**. In one approach, participants

[5] A universal version of this task was not used in this research; using the universal version did not effect the ability to answer the research question about the feasibility of this research, but the universal version was not possible since the date setting for the UUT was only available in an initial setup (when power is first applied to the unit) or as part of the installation procedures (covered in the installation manual). It is not covered in the user's manual. This situation will possibly confound comparisons between different units.

could be provided with a period of time to become familiar with the manual, though this does not ensure that they read or understand the manual and introduces a new variable into testing. As an alternative, participants could be provided access to the manual (as was originally proposed in this test), but the maximum time to complete the entire test should be based on the use of the user's manual. This method would introduce a question of the validity of the test since the usability of the manual would also be included and the tendency for people to refer to a manual is a random variable in the test. A third alternative procedure, and the simplest though most stringent on the design, is to require that participants be able to perform tasks without using the manual.

d) Summative Score versus Individual Scores

The question has been proposed as to whether the unit under test should have a single score from testing or should have individual scores per task. Each task in testing can be considered a single event of equal weight with all other tasks. This may not have face validity since some tests may be more critical to system operation than others. Regardless, whether the tests are weighted by an additional factor or equally weighted in the summative analysis, care should be taken into finding a single summative metric. For example, a unit where 50% of the population cannot perform one of the tasks versus a unit where 25% of the population failed to perform each of 2 tasks could show equal performance in a test if the summative score is the total number of successes versus the total number of failures across all tasks. For this reason, is recommended that the unit must meet the pass/fail criterion for each task in order to pass the test.