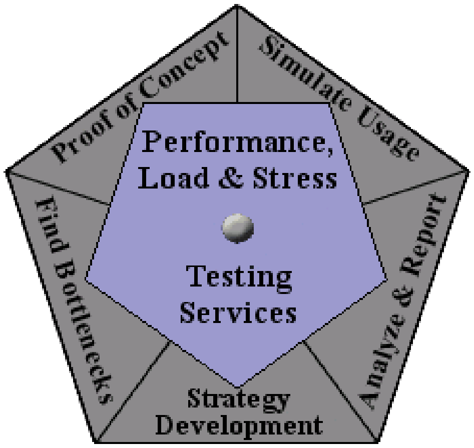
PerfTestPlus
Better Testing... Better Results

User Experience, Not Metrics

by:

R. Scott Barber



Part 7: Consolidating Test Results

"You've been running this test for weeks and sending me charts almost every day, but what does it all *mean*?!?" If your experience with performance testing is anything like mine, you've had someone say that to you at least once (in my case, several times). Managers and stakeholders need more than just the results from various tests — they need conclusions, and consolidated data that supports those conclusions. Later articles will address the topic of drawing conclusions from test results, but now I want to finish laying the groundwork by discussing the consolidation of results from multiple executions of identical tests.

This is the seventh article in the "User Experience, Not Metrics" series, which focuses on correlating customer satisfaction with your Web site application's performance as experienced by users. In Part 6 we discussed how to identify and account for outliers in your result sets. In the process of handling those outliers, you learned how to transfer the results from your timers into Microsoft Excel and duplicate the Response vs. Time scatter graph and the Performance Report Output graph from TestManager. This article starts where Part 6 left off, by having you execute a few more identical test runs and consolidate the results in Excel tables and graphs. Doing the consolidation itself is actually a simple matter; what's more involved is determining if test results can be consolidated.

Before reading this article, you should have read and worked through Part 6. Once again, there will be some statistical math involved, but as we've done previously, we'll take a commonsense approach and let Excel do the heavy math for us. This article is intended for all levels of TestStudio users and may also prove useful to managers of projects where performance testing will occur.

Why Consolidate Results?

While it isn't strictly necessary to consolidate results, I've found it to be much easier to show people patterns in results when those results are consolidated into one or two graphs rather than distributed over dozens. And sometimes you'll want to put the results from several test executions together into a single report to gain datapoints for statistical significance. For example, I've often run into situations where I'm restricted to running performance tests during one hour in the middle of the night. A single one-hour test doesn't provide much data to draw conclusive results from, but executing the exact same test during the same hour of the day on five consecutive days may provide enough data to draw conclusions from if I consolidate the results.

Determining If Test Results Can Be Consolidated

In order to be consolidated, test results must meet certain criteria. First, the text executions must be identical, and second, the test results must be statistically equivalent. I'll say more about each of these criteria, particularly about how to tell if the second criterion is met.

Are the Test Executions Identical?

For a series of test executions to be identical, both the test itself and the test environment must be identical.

The test itself. For tests to be identical, the same suite must be executed with the same parameters for the same number of users, though at different times. In earlier articles I've shown you how to develop test scripts with a certain amount of randomness built in, such as delay times randomly selected within a specified range, random navigational choices, or random data pulled from a datapool. Introducing this kind of randomness doesn't make one test significantly different from another test. On the other hand, increasing or decreasing the number of virtual testers, increasing or decreasing the number of iterations, adding or removing test scripts, or suppressing timing delays would make a test different enough to eliminate it from consideration for consolidation, because these changes would lead you to expect different results. The key to whether tests are identical is simply whether the test results *should* be the same.

The test environment. The test environment must be identical for test results to be consolidated. This means that the time of day, the external load on the system, the total volume of network traffic, the build of the application, the location and configuration of the load generation environment, and so forth must be the same for each test execution.

Are the Test Results Statistically Equivalent?

If test executions are identical, we still need to ask whether the results are statistically equivalent to see if they can be consolidated. Mathematically calculating statistical significance, equivalence, or correlation between samples is well beyond the scope of this article, so instead I'm going to show you how to create a graph comparing individual test results that will let you determine for yourself if the test results are equivalent. If you're interested in a mathematical approach involving Chi-squared, t-test, or p-value methods, you can find information on the StatSoft Inc. site.

In support of the commonsense approach described below, I want to share with you this excerpt from the StatSoft discussion on the topic:

"There is no way to avoid arbitrariness in the final decision as to what level of significance will be treated as really 'significant.' That is, the selection of some level of significance, up to which the results will be rejected as invalid, is arbitrary. In practice, the final decision usually depends on whether the outcome was predicted a priori or only found post hoc in the course of many analyses and comparisons performed on the data set, on the total amount of consistent supportive evidence in the entire data set, and on 'traditions' existing in the particular area of research. . . . But remember that those classifications represent nothing else but arbitrary conventions that are only informally based on general research experience."

With that in mind, and recognizing that there really aren't many "traditions" about statistical significance when dealing with response time over the Internet, we're going to create a graph like Figure 1 to simplify the comparison of individual test results. Figure 1 is a summary of the results of five executions of the same performance test whose results we charted in Part 6, a performance test consisting of 100 measurements against the Noblestar.com Web site taken by each of two timers (Home Page and Page1).

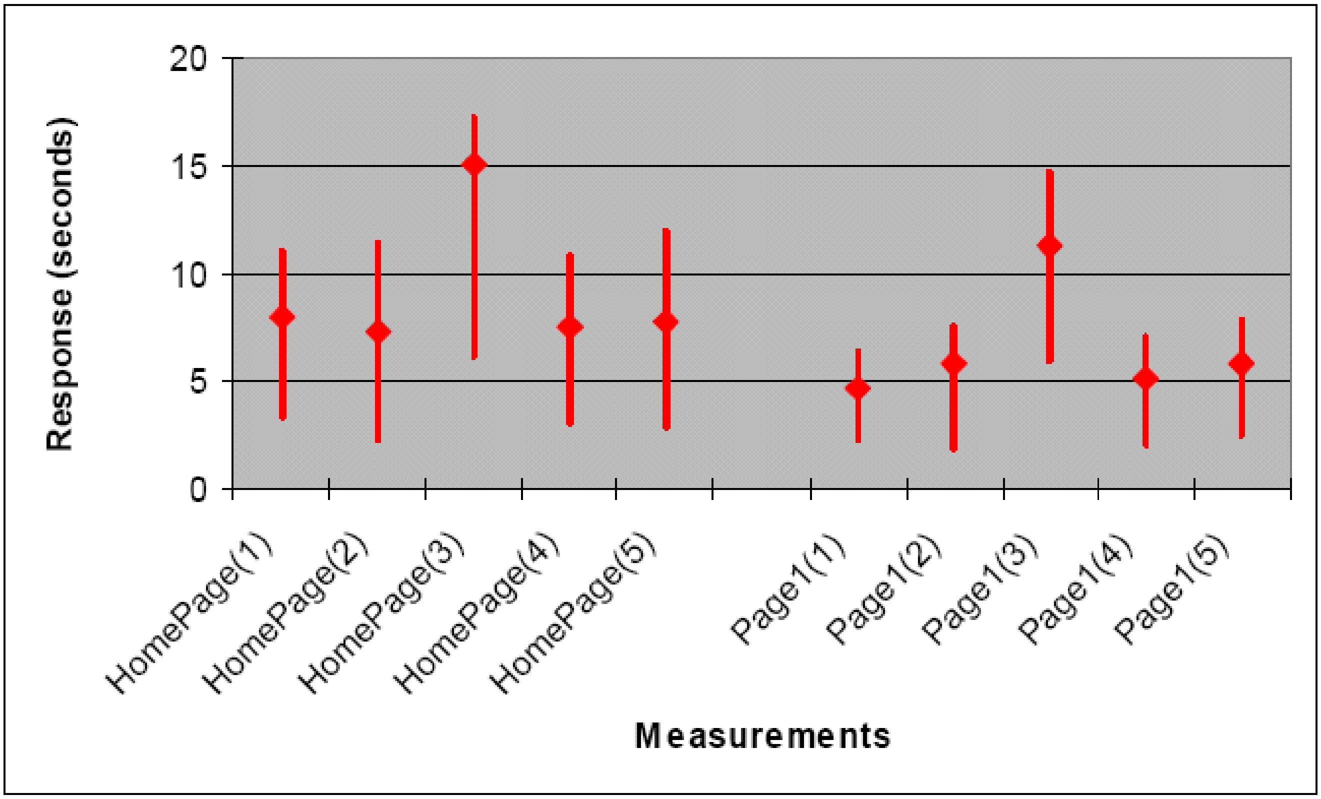


Figure 1: Results comparison graph

You may recognize this graph as an adaptation of the standard Excel stock chart. As you can see, the graph has a vertical line for each execution of the performance test for each of the timers. This makes it easy to compare the different executions of not just the performance test as a whole but also load times for individual pages. The bottom point of each red line is the minimum value, the top point of the line is the maximum value, and, in this case, the diamond marks the 95th percentile value.

When we look at the chart as a whole, we notice that the test executions labeled (1), (2), (4), and (5) show virtually identical results, but the results of the test execution labeled (3) lie noticeably higher on the chart. This is an example of a test execution anomaly. Test execution (3) obviously didn't produce results that are statistically similar to those of test executions (1), (2), (4), and (5).

Results won't often be this obviously statistically different. While there's no hard-and-fast rule about how to decide which results are statistically similar, I recommend that you compare results from at least five test executions and apply these rules of thumb to help you determine if test results are similar enough to be consolidated:

- If more than 20% (or one out of five) of the test execution results appear *not* to be similar to the rest, something is generally wrong with either the test environment, the application, or the test itself.

- If a 95th percentile value for any test execution is greater than the maximum or less than the minimum value for any of the other test executions, it's not statistically similar.

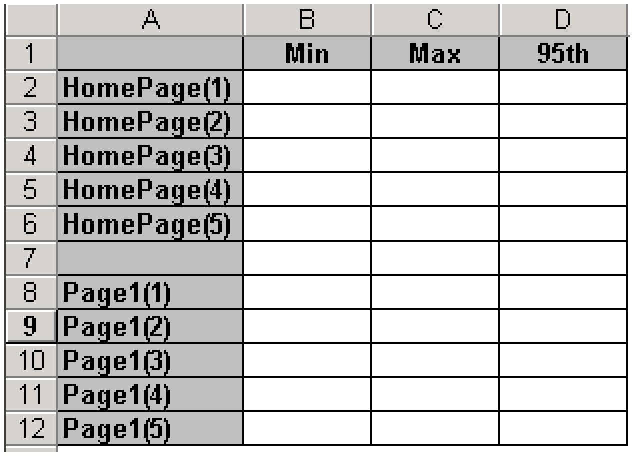
- If every page/timer result in a test execution is noticeably higher or lower on the chart than the results of all the rest of the test executions, it's not statistically similar.

- If a single page/timer result in a test execution is noticeably higher or lower on the chart than all the rest of the test execution results, but the results for all the rest of the pages/timers in that test execution are not, the test executions are probably statistically similar.

To create the graph, I began by executing the performance test whose results we charted in Part 6 five times and adding the results to the Excel worksheet I first created there, handling outliers as described there. To create your own results comparison graph, add the results of executing your own performance test at least five times to your Excel worksheet and then follow these steps:

- Create a new worksheet by choosing Insert > Worksheet from the Excel menu bar. You can move the new worksheet by selecting the tab on the bottom and dragging the tab to the position you desire. You can also rename the tab by double-clicking on it and typing the new name, in this case "Compare Results."

- On the new worksheet, create a table with columns for the minimum, maximum, and 95 th percentile values and with the names of the pages/timers down the left side, as shown in Figure 2. It's important to leave a blank line in between each group of pages/timers.

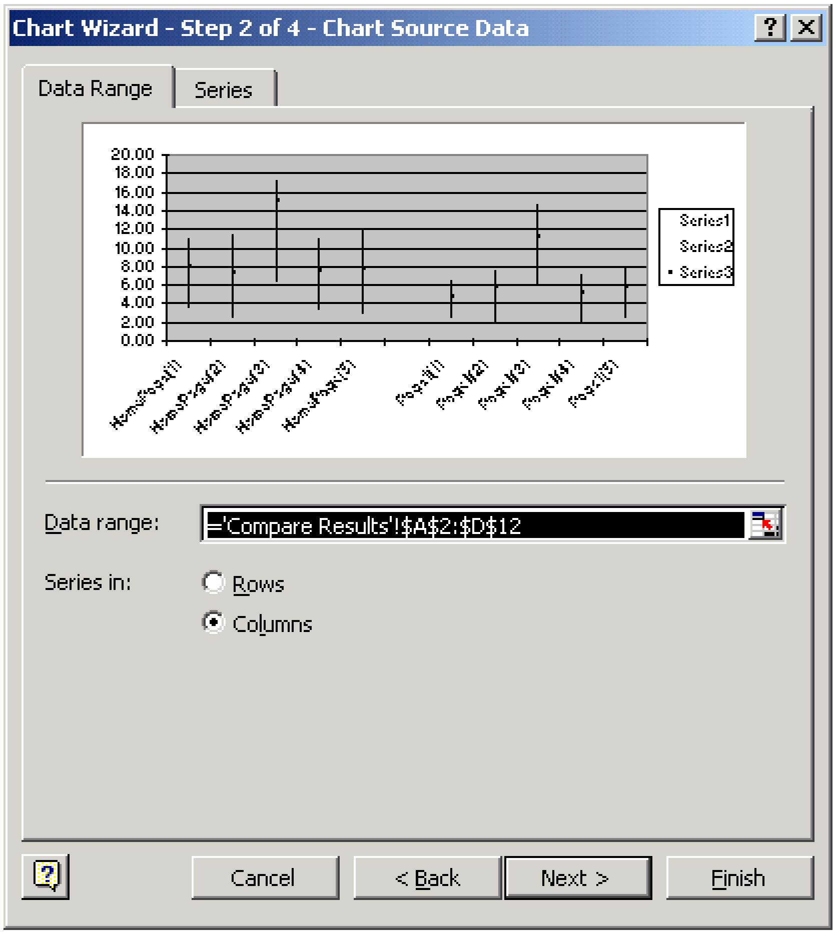


	A	B	C	D
1		Min	Max	95th
2	HomePage(1)			
3	HomePage(2)			
4	HomePage(3)			
5	HomePage(4)			
6	HomePage(5)			
7				
8	Page1(1)			
9	Page1(2)			
10	Page1(3)			
11	Page1(4)			
12	Page1(5)			

Figure 2: Results comparison table format

- Populate the table with data values. You can do this by copying and pasting the values from the previous worksheet, typing in the values by hand, or linking the cells to the values in the previous worksheet by typing "=" in a cell, navigating to the cell containing the value in the other worksheet, and clicking Enter.

- Highlight your table, choose Insert > Chart… from the menu bar, and on the Standard Types tab, under "Chart types," select Stock. Once again, there are many options and configuration possibilities for this chart that you can explore on your own. For now, just be sure that the Columns option is selected on the Data Range tab, as shown in Figure 3.



Data range: ='Compare Results'!$A$2:$D$12

Series in: ○ Rows ● Columns

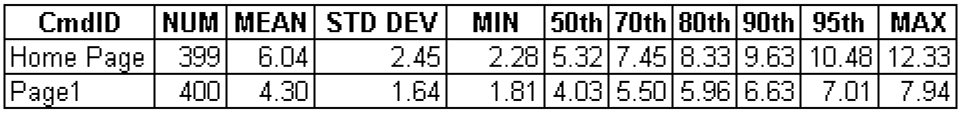
Figure 3: Selecting the Columns option on the Data Range tab

The resulting chart should make it easy for you to compare the results of multiple executions of your performance test to determine whether they're statistically equivalent. Assuming that they are, you can proceed to consolidate the results.

Creating a Performance Report with Consolidated Results

Creating our standard performance report output table and chart with consolidated results is a simple process, although to arrive at our consolidated performance response statistics, we can't simply average the results from our test executions together. Rather, we need to actually recalculate the statistics. Once again, instead of trying to figure out complex formulas to generate weighted averages or to find absolute percentiles across multiple result sets, we'll take a commonsense approach.

1. Create a new worksheet by choosing Insert > Worksheet from the Excel menu bar.
2. For each of the test executions to be consolidated, copy and paste columns A, B, and C from the performance response worksheets you've previously created in Excel to handle outliers if you copied in all of the timer information, or the three columns (Cmd ID, Ending TS, and Response) from the Response vs. Time Report Output in TestManager. Paste each successive set of results below the previous one and then sort by column A.
3. Follow the instructions under "Recalculating Performance Report Values" in Part 6 to create the consolidated performance report output table, as shown in Figure 4.



CmdID	NUM	MEAN	STD DEV	MIN	50th	70th	80th	90th	95th	MAX
Home Page	399	6.04	2.45	2.28	5.32	7.45	8.33	9.63	10.48	12.33
Page1	400	4.30	1.64	1.81	4.03	5.50	5.96	6.63	7.01	7.94

Figure 4: Consolidated performance report output table

4. Once the table is populated, simply follow the instructions under "Recreating the Performance Response Chart" in Part 6 to create the consolidated performance report output chart. I used the performance results output table without the NUM and STD DEV columns to create the chart shown in Figure 5. You may choose any combination of the columns to include in this chart. The instructions in Part 6 describe how to make this chart without the MEAN column. It is entirely your choice. I recommend always charting MIN, MAX and either the 90th or 95th percentile measurements.

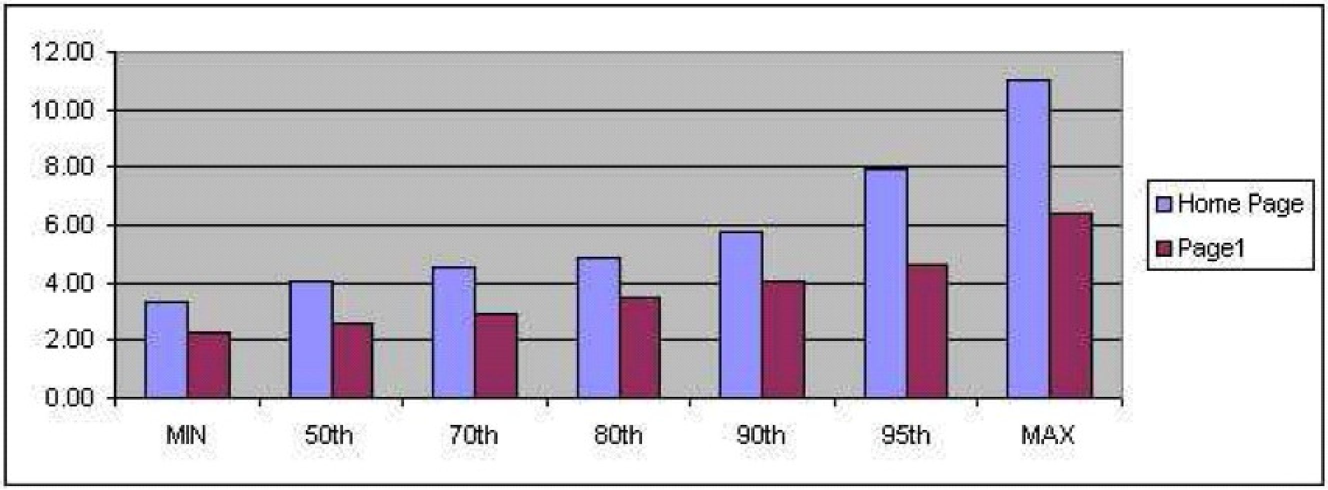


Figure 5: Consolidated performance report output chart

Now You Try It

If you want to try the approach to consolidating results that I suggest here, begin with the exercise you completed for Part 6, execute the same test script four more times, and follow the steps above to consolidate your results as appropriate.

Summing It Up

The key point of this article is that only results from identical test executions that are statistically similar can be consolidated into performance report output tables and charts. Once you've constructed a chart to determine if results are statistically similar, the process of creating consolidated tables and charts is simple. The following three articles in this series will discuss how to create various types of tests and reports that will help in drawing conclusions and presenting them graphically.

Related Resources

1) An online statistical textbook is available at StatSoft Inc.

Acknowledgments

- The original version of this article was written on commission for IBM Rational and can be found on the IBM DeveloperWorks web site

About the Author

Scott Barber is the CTO of PerfTestPlus (www.PerfTestPlus.com) and Co-Founder of the Workshop on Performance and Reliability (WOPR – www.performance-workshop.org). Scott's particular specialties are testing and analyzing performance for complex systems, developing customized testing methodologies, testing embedded systems, testing biometric identification and security systems, group facilitation and authoring instructional or educational materials. In recognition of his standing as a thought leading performance tester, Scott was invited to be a monthly columnist for Software Test and Performance Magazine in addition to his regular contributions to this and other top software testing print and on-line publications, is regularly invited to participate in industry advancing professional workshops and to present at a wide variety of software development and testing venues. His presentations are well received by industry and academic conferences, college classes, local user groups and individual corporations. Scott is active in his personal mission of improving the state of performance testing across the industry by collaborating with other industry authors, thought leaders and expert practitioners as well as volunteering his time to establish and grow industry organizations. His tireless dedication to the advancement of software testing in general and specifically performance testing is often referred to as a hobby in addition to a job due to the enjoyment he gains from his efforts.

About PerfTestPlus

PerfTestPlus was founded on the concept of making software testing industry expertise and thought-leadership available to organizations, large and small, who want to push their testing beyond "state-of-the-practice" to "state-of-the-art." Our founders are dedicated to delivering expert level software-testing-related services in a manner that is both ethical and cost-effective. PerfTestPlus enables individual experts to deliver expert-level services to clients who value true expertise. Rather than trying to find individuals to fit some pre-determined expertise or service offering, PerfTestPlus builds its services around the expertise of its employees. What this means to you is that when you hire an analyst, trainer, mentor or consultant through PerfTestPlus, what you get is someone who is passionate about what you have hired them to do, someone who considers that task to be their specialty, someone who is willing to stake their personal reputation on the quality of their work - not just the reputation of a distant and "faceless" company.