

Beginner's Guide to Measuring Educational Outcomes in CEhp

How to Analyze Your Baseline/Post-Activity Change Data

Part 2: Baseline/Post Rating Scale (Ordinal) Questions

By Erik D. Brady, PhD, CHCP, EDBPHD Consulting, and Derek T. Dietze, MA, FACEHP, CHCP, Improve CME, LLC

In our last article, we addressed how to analyze results from multiple choice questions asked both before and after a CEhp activity. Another common type of baseline/post-activity question involves the use of a rating scale to assess changes in confidence, agreement or frequency of use. Again, these data are collected at the time of the activity on paper forms or through an Audience Response System (ARS). Summarizing the results for each baseline/post-activity question and calculating a “*P* value” for the change in ratings can provide insights into the effectiveness of your CEhp activity. It can enhance the credibility of your outcomes reports and provide a foundation for improving future activities. Finally, with appropriate goal statements within your mission statements, these types of measures can be a powerful way to analyze your overall CEhp program.

Scope of this Article

This article focuses on providing step-by-step directions on how to calculate a *P* value for baseline/post-activity rating scale questions. The results data from these questions are considered “ordinal” data—they have an order (i.e., lowest to highest). We have highlighted two cases: The first addresses a situation in which you have collected paired data, and the second addresses a case in which you have collected unpaired data (see “Basic Concepts of Data Sets” in the [September 2015](#) issue of the *Almanac*). You will also find it helpful to review the concepts of data that are “normally” or “not normally distributed,” how to choose a “parametric” or “non-parametric” statistical test (see “Distribution and Variation in Data Sets” in the [October 2015](#) issue of the *Almanac*), and the definition of a “*P* value” (see “How to Analyze Your Baseline/Post Activity Change Data Part 1: Baseline/Post Multiple Choice Questions in the December 2015 *Almanac*”).

Working with Paired Data

As with multiple choice items, having a data set in which the

Table 1. A set of paired data from a typical educational activity; *n* = number of learners responding to both the baseline and post instance of the question

Question #	<i>n</i>	Baseline Average (Current Use)	Post Average (Planned Use)	Change (Δ)
1	21	2.57	5.33	+2.76
2	48	5.63	5.98	+0.35
3	56	5.43	5.70	+0.27
4	32	3.47	5.91	+2.44
5	17	5.59	5.65	+0.06

responses to rating scale items are assigned to specific individuals is definitely a preferred situation, as it allows us to work with a paired data set. As with unpaired data, the first step is to calculate the group baseline average and the group post average to determine the “delta,” or “change,” for the group being considered. Note that the symbol for delta is Δ . At that point, we run a distinct test to calculate the *P* value. As we have done for multiple choice items (in the previous article) and with unpaired ordinal data (later in this article), we show the best way to describe the calculation using an example.

Case 1: Paired Baseline/Post-Activity Rating Scale Question Data (Parametric)

A data set for your online educational activity that was recently conducted had five rating scale questions that were asked after the delivery of content to assess changes in intent (competence) for specific practice strategies that were supported by the activity content. A seven-point semantic differential scale was used (descriptive words only at each end of the scale, but not for the values two through six) in each case: 1 = No Use and 7 = Extensive Use. Learners were able to respond to question items as they desired, but the data analysis was restricted to only those who offered a response to both a baseline and a post question for each question item. The resulting data was found across the activity, as shown in [Table 1](#). All changes

appear positive, and when you share them with the course director, he/she indicates a desire to understand the significance of these findings.

Step 1: Access a Statistical Computation Tool

In order to determine a *P* value for paired ordinal data, several tests are available. One challenge of working with ordinal data is that you need to understand whether or not your data are parametric (i.e., shaped like a Bell curve) or non-parametric (i.e., not shaped like a Bell curve). A t-test, which we describe below, is a parametric test.

When we graph our ordinal data, particularly when we ask learners to rate something (e.g., confidence or intent-to-use practice strategy), it is not atypical to find that our data resembles a Bell curve, or would resemble a Bell-shaped curve if the data were extrapolated. If that is the case, then using a paired t-test is totally appropriate.¹ If the data looks flat or in some way does not resemble a Bell curve, then we are in a position where we need to use a non-parametric test. In the case of paired ordinal data, the Wilcoxon signed-rank test is the most appropriate test to use.¹ We will direct readers to easy online tools for both the t-test and the Wilcoxon test, and you can use a free online tool from [Social Science Statistics](#).

An easy tool for the paired t-test can be found at [GraphPad](#). As with all the tools that we refer to, access is free and available online. In order to access the appropriate tests to analyze the data found in Table 1, access the website, select “Continuous Data” on the first screen, click “Continue,” then select “t-test to compare two means” and click “Continue,” or go directly to the [Web page](#). **Figure 1** shows the screen that will appear to assist in your calculation of a *P* value using paired ordinal data.

Step 2: Organize Your Data and Execute Calculation

In general, it’s typically easiest to select the second radio button, “Enter or paste up to 2000 rows,” unless you have a very small data set and prefer to manually key in the data. You will need to ensure that each row in your data set represents an individual learner. “Group One” is the baseline cohort, so all baseline data should be copied and pasted into the first column shown in Figure 1. “Group Two” is the post cohort. The tool does allow you to replace the labels with “Baseline” and “Post” if you so choose. Once you add your data to the Group One and Group Two columns, select “Paired T-test” under “3. Choose a Test.” Click “Calculate Now” under “4. View the Results” to return the *P* value for your paired data set, as well as several other pieces of information. An example of the returned data set is shown in **Figure 2** for Question 1.

Figure 1. T-test input screen on GraphPad

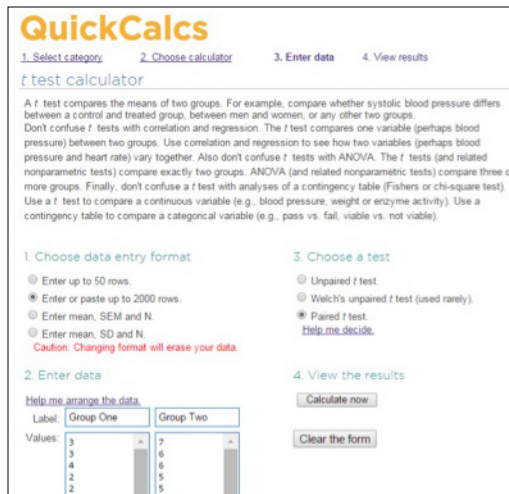


Figure 2. A sample calculation in GraphPad with the results page shown



Table 2. Addition of calculated P values to assess significance of change percentages (Δ).

Question #	n	Baseline Average (Current Use)	Post Average (Planned Use)	Change (Δ)	P value
1	21	2.57	5.33	+2.76	<.0001
2	48	5.63	5.98	+0.35	.0022
3	56	5.43	5.70	+0.27	.0799
4	32	3.47	5.91	+2.44	<.0001
5	17	5.59	5.65	+0.06	.5795

The key value is the “two-tailed *P* value,” determined as < 0.0001 for the example Question 1, which is shown framed in a red box in Figure 2. When the t-test is performed for all five example questions, we can add *P* values to our original table, shown as in **Table 2**.

Step 3: Analyze Your Change

While your first glance at the data showed positive change

on all items, when you consider the *P* values, you find that the significance of the calculated change from baseline to post (deltas) are greatly varied. For example, for Question 1, you see only 21 total matched responses – an “n” that you might expect to push the boundaries of a significance calculation. Clearly, this is a highly significant finding, even with a fairly low number of responses. Question 4 is similar; baseline and post means are higher and are across a larger number of respondents than Question 1, and we find a similar *P* value.

Question 2 has a larger number of responses, but the change from baseline to post is smaller than what was found for Question 1. As was noted in the last article in the series, this *P* value (0.0022) is still lower than the threshold used by most to qualify for significance at the 95 percent confidence level (< 0.05).

Questions 3 and 5 are also included here for specific reasons. With Question 3, the number of responses is 56, which is the largest n in the data set, and yet, we find a *P* value of 0.0799, which is considered not statistically significant, but may be described as trending toward significance. A higher number of respondents may lead to a *P* value < 0.05. Question 5 has only 17 responses, barely allowing for validity of the calculation. The *P* value for Question 5 is 0.5795, also an insignificant finding. The core message is that simply increasing the number of respondents doesn’t always lead to a significant finding, but this must be considered on a case-by-case basis.

Case 2: Paired Baseline/Post-Activity Rating Scale Question Data (Non-parametric)

If shortly after completing your analysis from the previous case you take a look at a bar chart representation of your data and find that it does not appear to be parametric, you may question the *P* values that you calculated using the t-test. In order to feel greater confidence, you decide to re-analyze your original data set using a non-parametric test. The Wilcoxon signed-rank test is an appropriate test for paired ordinal data that are not normally distributed.¹

Step 1: Access a Statistical Computation Tool
An easy tool for the Wilcoxon signed-rank test can be found on the [Social Science Statistics website](#). **Figure 3** shows the screen that will appear to assist in your calculation of a *P* value using paired data.

Step 2: Organize Your Data and Execute Calculation
As with the tool available on GraphPad, the Wilcoxon signed-rank tool requires data to be formatted into baseline (Treat-

Figure 3. Wilcoxon signed-rank test input screen on Social Science Statistics

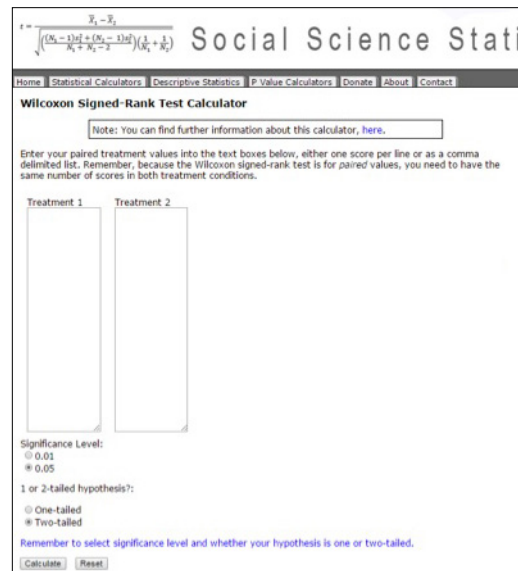
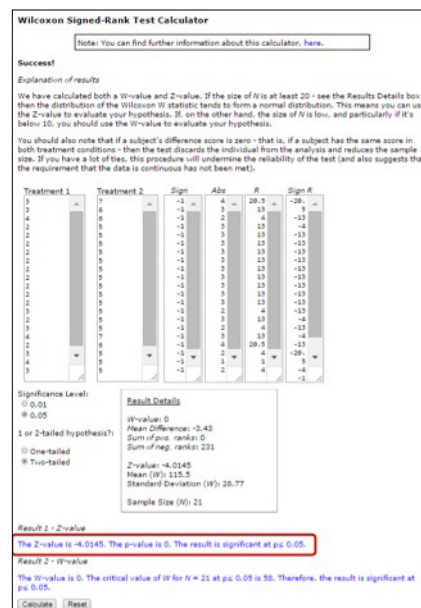


Figure 4. A sample Wilcoxon calculation with the results page shown



ment 1) and post (Treatment 2) groups. Data are pasted or hand-keyed into each box. Once “Significance Level: 0.05” and “Two-tailed” hypothesis have been indicated, click “Calculate.”

The key value is the “Z-value,” determined as “0” for the example Question 1, which is shown framed in a red box in **Figure 4**. It is not atypical to show “0” as “<0 .0001” to show a consistent precision in the analysis of *P* values. When the Wilcoxon signed-rank test is performed for all five sample questions, we can add *P* values to our original table, as shown in **Table 3**.

Step 3: Analyze Your Change

In general, the interpretation of your results is the same regardless of whether you select a parametric or a non-parametric test of significance for ordinal data. Comparing the calculated *P* values reveals minimal changes for findings that are highly significant. For example, for Questions 1 and 4, both parametric and non-parametric tests result in a *P* value <0.0001 regardless of the chosen test. Question 2 also falls into the category of significant with either test.

Question 3, however, presents an interesting finding, in that the parametric test suggested near significance (*P* = 0.0799), and the non-parametric test (*P* = 0.0173) falls into the range of significance. The non-parametric test would allow you to call this significant, because we found that our data was not Bell shaped in this case.

Question 5 has only 17 responses, barely allowing for validity of the calculation with the t-test (*P* = 0.5795). The Wilcoxon signed-rank test cannot be performed for samples that are this small, having a requirement of 17 non-zero differences in order to be valid.

Case 3: Unpaired Baseline/Post-activity Rating Scale Question Data (Parametric)

From your hospital grand rounds CME activity, you collected participants' answers to four confidence rating scale baseline (pre) questions before the activity and the same questions post-activity (i.e., please rate your confidence in your ability to XYZ: 1 = Not confident at all, 2 = Not very confident, 3 = Somewhat confident, 4 = Very confident, 5 = Extremely confident). You have a stack of completed baseline questionnaires and a stack of post questionnaires, and there are no names or email addresses (unique identifiers) on the questionnaires, so you cannot match them. Also, you have 38 completed baseline questionnaires and 31 completed post questionnaires, because some participants left early and did not complete the post questionnaire. How do you determine if there was a statistically significant increase in ratings for each confidence question?

Step 1: Enter Your Data into Excel

Table 4 shows what your data should look like in Excel after initial data entry. Due to space limitations in this article, we are only showing results from the first eight completed baseline questionnaires and the first five completed post questionnaires.

All changes appear positive, and when you share them with the course director, he/she indicates a desire to understand the significance of these findings.

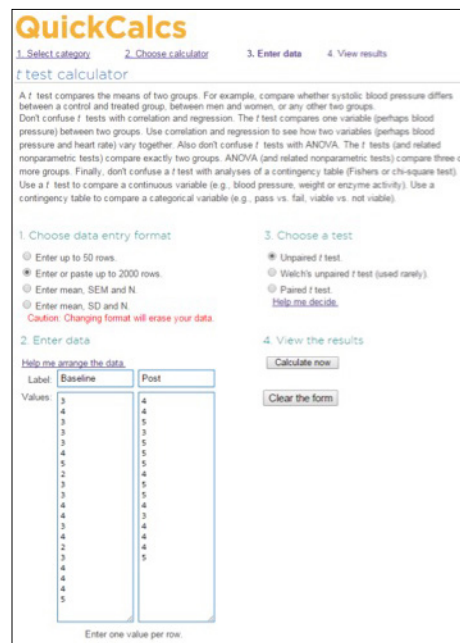
Table 3. Addition of calculated P values to assess significance of change percentages (Δ)

Question #	n	Baseline Average (Current Use)	Post Average (Planned Use)	Change (Δ)	P value
1	21	2.57	5.33	+2.76	<.0001
2	48	5.63	5.98	+0.35	.0050
3	56	5.43	5.70	+0.27	.0173
4	32	3.47	5.91	+2.44	<.0001
5	17	5.59	5.65	+0.06	N/A

Table 4. Unpaired Data Entry in Excel

Baseline or Post	Q1	Q2	Q3	Q4
Baseline	3	1	4	3
Baseline	4	2	5	3
Baseline	4	3	5	3
Baseline	3	2	5	5
Baseline	3	3	5	3
Baseline	5	4	4	3
Baseline	4	3	4	3
Baseline	3	3	3	4
Post	3	2	5	4
Post	4	2	5	5
Post	4	3	4	5
Post	5	3	5	4
Post	5	4	5	5

Figure 5. Unpaired t-test data entry screen in GraphPad



Initially you make the assumption that your data are normally distributed (Bell-shaped). The appropriate test for normally distributed unpaired ordinal data is the unpaired t-test, a parametric test.¹

Step 2: Access a Statistical Computation Tool and Enter Your Data

As with the paired data analysis, the t-test is still an appropriate test for unpaired data, so you used the same tool found at [GraphPad](#). Under “1. Choose Data Entry Format,” select “Enter or Paste up to 2000 Rows,” and under “3. Choose a Test” select “Unpaired T-test” as shown in **Figure 5**. Label the first column “Baseline” and the second column “Post,” then for confidence question one (Q1) in your Excel file, copy and paste the raw baseline data into the first column, and the raw post data into the second column. For space reasons, not all data are shown in the figure, but for Q1 in our example case, you have 38 data points in the first column and 31 in the second.

Step 3: Calculate the P Value for Q1, Repeat for Q2-Q4 Under “4. View the Results,” click “Calculate Now.” **Figure 6** shows the results with a two-tailed P value of 0.0036 (see the red box), indicating that there was a statistically significant increase in confidence ratings for Question 1 from baseline (3.50/5) to post (4.31/5), $P = 0.0036$, baseline $n = 20$, post $n = 16$, unpaired t-test. Repeat the same procedure to obtain P values for questions two through four, and summarize your results in a table, much like that shown for the paired data cases above.

Case 4: Unpaired Baseline/Post Rating Scale Question Data (Non-parametric)

Shortly after completing your analysis, you find that your data does not appear to be parametric, which makes you question the P values that you determined using the unpaired t-test. In order to feel greater confidence, you decide to re-analyze your original data set using a non-parametric test. The most appropriate test for unpaired ordinal data that are not normally distributed is the Mann-Whitney U test.¹

Step 1: Access a Statistical Computation Tool and Enter Your Data

An easy tool for the Mann-Whitney U test can be found at [Social Science Statistics](#). **Figure 7** shows the screen that will appear to assist in your calculation of a P value using unpaired ordinal data. The Mann-Whitney U test tool requires data to be formatted into baseline (Population 1) and post (Population 2) groups. Data are pasted or hand-keyed into each box.

Step 2: Calculate the P Value for Q1, Repeat for Q2-Q4 With “Significance Level: 0.05” and “Two-tailed” hypo-

Figure 6. A sample calculation for the unpaired t-test in GraphPad with the results page shown

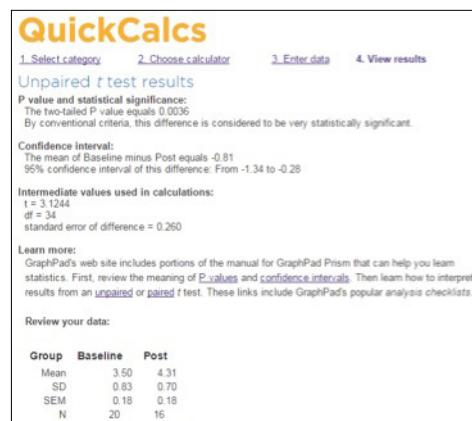


Figure 7. Mann-Whitney U-test input screen on Social Science Statistics

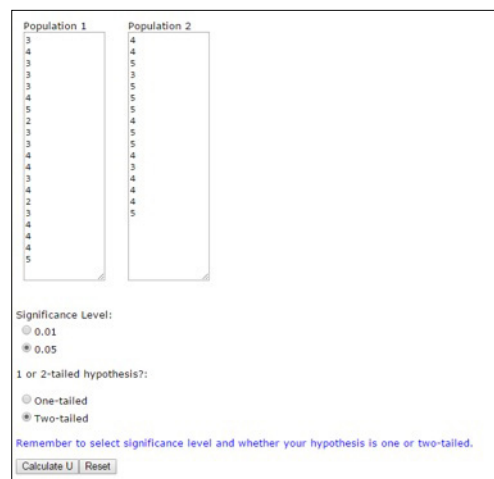


Figure 8. A sample Mann-Whitney U test calculation with the results page shown




esis indicated, click “Calculate.” The results are shown in **Figure 8**. A P value of 0.00854 is shown in the red box, indicating that there was a statistically significant increase in confidence ratings for question one from baseline (3.50/5) to post (4.31/5), $P = 0.0085$, baseline $n = 20$, post $n = 16$, Mann-Whitney U test. Repeat the same procedure to obtain P values for questions two through four, and summarize your results in a table. Review your results as described in the paired data case. An alternative online tool for the Mann-Whitney U test can be found at [Vassar Stats](#) by clicking on “Ordinal Data” on the menu at the left of the page and then on “Mann-Whitney Test.”

Limitations

Some of the same limitations that we described in the previous article on multiple choice items continue to hold true with ordinal data, and we restate them here. For unpaired data sets, there are definite limitations when the n of the baseline and post groups are highly varied. In that case, it’s possible that the two groups may not be an accurate reflection of each other. For example, consider a scenario where the baseline group has 150 responses and the post group has 30 responses. In addition, the 30 post responses are all members of your target audience, but the 150 baseline response are a mix of target audience and non-target audience.

It’s possible that the calculated delta and P values may be less valid than your calculations would suggest. This is one of the rationales for using paired data whenever possible.

Open-access online statistical test tools can be used to calculate P values for your paired or unpaired ordinal data (i.e., rating scale data) that you collect to assess the effectiveness of your CEhp activities. For paired ordinal data, the paired t -test is best when the data are normally distributed, and the Wilcoxon signed-ranks test is best when the data are not normally distributed. For unpaired ordinal data, the unpaired t -test is best for normally distributed data, and the Mann-Whitney U test is best when the data are not normally distributed. 

Reference

1. How to Choose a Statistical Test. <http://www.graphpad.com/support/faqid/1790/> Accessed 12/12/15.

Resources

2. Jason Olivieri, CMEPalooza, Statistical Analysis in CME Outcomes, <http://cmepalooza.com/march21/statistical-analysis-in-cme-outcomes-olivieri/>
3. Erik D. Brady, PhD, CHCP, CMEPalooza “Excel”lent Tricks for the Non-Expert: Exploring the Beauty of the Cells. <https://www.youtube.com/watch?v=11I75UrlqxE>



REACH
HEALTHCARE PROFESSIONALS
FAR & WIDE

We get your message delivered.
You get responses.

CONTACT US AT 800.MED.LIST OR SALES@MMSLISTS.COM

 **mms**
message delivered.
mmslists.com • 800.MED.LIST