Beginner's Guide to Measuring Educational Outcomes in CEhp

How to Analyze Your Baseline, Post-Activity Change Data

Part 1: Baseline, Post-Activity Multiple-Choice Questions

By Erik D. Brady, PhD, CHCP, Wake Forest Baptist Medical Center; and Derek T. Dietze, MA, FACEhp, CHCP, Improve CME, LLC

This article addresses ACEhp National Learning Competency:

- Competency Area 3.1: Measuring the Performance of Activities and the Overall Program. Use evaluation and outcomes data … (C) Analyzing assessment data in order to draw conclusions about the effectiveness of the activity/intervention based on expected results.

One of the most common types of outcomes data that CEhp professionals can work with comes from multiple-choice knowledge and competence questions asked both before and after a CEhp activity. These data are typically collected at the time of the activity via paper forms, online or an Audience Response System (ARS). Summarizing the results for each baseline to post-activity question and calculating a "*P* value" for the change in number of correct answers from baseline to post-activity can provide insights into the effectiveness of your CEhp activity. It can also enhance the credibility of your outcomes reports, and provide a foundation for improving future activities.

This article focuses on providing a working definition of *P* value and provides step-by-step directions on how to calculate a *P* value for baseline to post-activity multiple-choice knowledge and competence questions. Two cases are highlighted: the first addresses collected non-paired data, and the second highlights collected paired data (for more information about paired and non-paired data, see the article "Basic Concepts of Data Sets," published in the September 2015 issue of the *Almanac*).

What is a "*P* value"?

A *P* value (the "*P*" means "probability") is generated from a test of statistical significance (a mathematical formula).[1] In the case of comparing baseline answers to posttest answers of multiple-choice questions, the *P* value indicates whether or not the before-to-after change in correct answers was statistically significant. Simply put, the *P* value represents the role that chance plays in your outcomes.

The calculation used for *P* value results in a value between 0 and 1 and can be interpretted.[2] In general, a *P* value of .05 or less represents the "gold standard" in scientific research, meaning that 95 percent of the time your findings are statistically significant. This means that there is only a 5 percent likelihood that a calculated change from baseline to post-activity would occur by chance alone if the same education were offered to additional learners of similar demographics.

Statistical significance does not necessarily mean practical significance. Only by considering context can you determine whether a difference is practically significant (that is, whether it requires action).[1]

In general, if there is an increase in correct answers from baseline to post-activity, you want to see a *P* value of 0.05 or lower in order to state in your outcomes report, "There was a statistically significant increase in correct answers from baseline to post."

- A small *P* value (typically ≤ 0.05) indicates strong evidence that the baseline to post change is real and is not due to chance. An *increase* in correct answers from baseline to post with a *P* value of ≤ 0.05 is a *positive* result — a statistically significant increase in correct answers. A *decrease* in correct answers from baseline to post, with a *P* value of ≤ 0.05 is a *negative* result — a statistically significant decrease in correct answers.
- A large *P* value (> 0.05) indicates weak evidence that the baseline to post change is real, and it is more likely due to chance. An increase in correct answers from baseline to post, with a P value >0.05 means that while more people answered correctly post than at

baseline, the increase was not statistically significant. Conversely, if there was a decrease in correct answers with a *P* value of >0.05, that decrease was not statistically significant or meaningful.

Sometimes analysts will refer to a "null hypothesis" and an "alternative hypothesis"[2] when conducting tests of statistical significance. In the context of baseline/post-activity multiple-choice questions, the null hypothesis is that *there is no difference* between correct answers baseline and post-activity. The statistical test determines if this null hypothesis is correct or not. If you get a *P* value of <0.05, then you reject the null hypothesis and accept the alternative hypothesis, which is that *there is a difference* between correct answers baseline and post.

Case: Unpaired Baseline/Post Multiple-Choice Question Data

In this case study, assume that from your hospital grand rounds CME activity you collected participants' answers to six multiple-choice baseline questions before the activity and the same questions post-activity. You have a stack of completed baseline questionnaires and a stack of post questionnaires, and there are no names 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 correct answers for each multiple- choice question?

Step 1: Enter your data into Excel.

Table 1 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. Also, we show only data for three of the six questions. Notice that beside each column where you have entered each participants' answer to a question (a, b, c, or d), you have "coded" their answer as either correct (1) or incorrect (0). Since the correct answer for question 1 is B, you have labeled the column "Q1CorrectB" to help with your coding.

Step 2: Summarize the number of correct and incorrect responses in a table.

The remainder of these steps focuses on question one results. You would repeat these steps for each of the six questions. After doing your data entry for all 38 baseline questionnaires and all 31 post questionnaires for baseline question 1, you count up the 25 participants who answered correctly and the 13 who answered incorrectly.

Table 1. Unpaired Data Entry and Correct Answer Coding

Pre or Post	Q1	Q1 Correct B	Q2	Q2 Correct A	Q3	Q3 Correct D
Pre	a	0	a	1	d	1
Pre	b	1	d	0	d	1
Pre	b	1	c	0	c	0
Pre	b	1	a	1	b	0
Pre	c	0	a	1	c	0
Pre	b	1	d	0	d	1
Pre	d	0	d	0	d	1
Pre	c	0	b	0	d	1
Post	b	1	a	1	d	1
Post	c	0	d	0	d	1
Post	b	1	a	1	c	0
Post	b	1	a	1	d	1
Post	b	1	c	0	d	1

Table 2. Question 1 Baseline/Post Correct/Incorrect Answers

	Correct	Incorrect
Pre	25	13
Post	27	4

Table 3. Blank 2x2 Contingency Table

	Outcome 1	Outcome 2
Group 1		
Group 2		

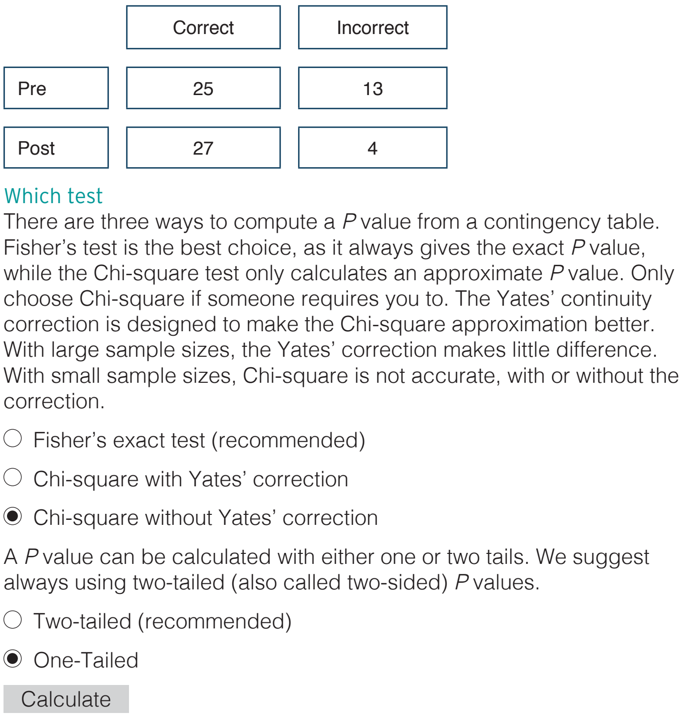
For post question one, 27 answered correctly and four incorrectly. Using this information, in Excel, create **Table 2**. Notice that you have used the count of correct/incorrect answers, not percentages.

Step 3: Enter results in an online tool to calculate the P value.

Proceed to a free online statistics tool to enter your data. While many are available, GraphPad is a simple one to use. **Table 3** shows a simple table (called a "2x2 contingency table") as shown on the Web page where you will enter your data. Type in and replace "Outcome 1" with "Correct," "Outcome 2" with "Incorrect," "Group 1" with "Pre" and "Group 2" with "Post." Then enter the data from the Excel table you created in Step 2.

What you entered should now look like **Figure 1**, and as shown, you select "Chi-square without Yates' correction" as the test you want completed, select "Two-tailed" and press "Calculate." If any numeric value you enter into the table (as shown in Figure 1) is five or less, it is recommended that you select "Fisher's Exact Test" under "Which Test," instead of the Chi-square test.

Figure 1. Completed Table and Selection of Test and Tails



	Correct	Incorrect
Pre	25	13
Post	27	4

Which test

There are three ways to compute a *P* value from a contingency table. Fisher's test is the best choice, as it always gives the exact *P* value, while the Chi-square test only calculates an approximate *P* value. Only choose Chi-square if someone requires you to. The Yates' continuity correction is designed to make the Chi-square approximation better. With large sample sizes, the Yates' correction makes little difference. With small sample sizes, Chi-square is not accurate, with or without the correction.

○ Fisher's exact test (recommended)

○ Chi-square with Yates' correction

● Chi-square without Yates' correction

A *P* value can be calculated with either one or two tails. We suggest always using two-tailed (also called two-sided) *P* values.

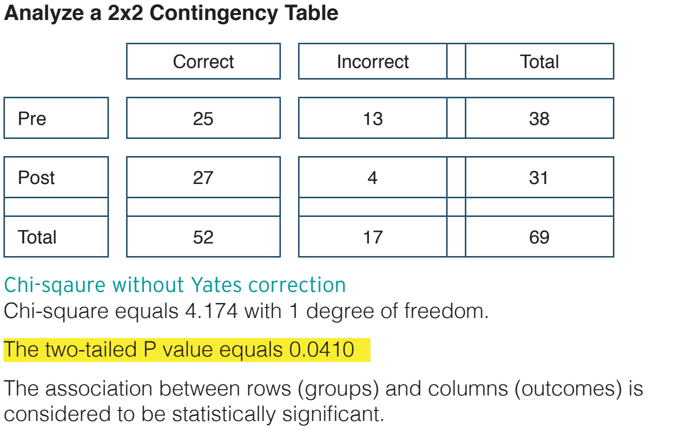
○ Two-tailed (recommended)

● One-Tailed

Step 4: Review results and create a significance statement.

The *P* value calculated using this method is 0.041, as shown in **Figure 2** highlighted in yellow. Thus, your statement regarding question one would be, "There was a statistically significant increase in correct answers from baseline to post (*P*=0.041, baseline n=38, post n=31, Chi-square test)."

Figure 2. *P* Value from Chi-square Test



Analyze a 2x2 Contingency Table

	Correct	Incorrect	Total
Pre	25	13	38
Post	27	4	31
Total	52	17	69

Chi-sqaure without Yates correction

Chi-square equals 4.174 with 1 degree of freedom.

The two-tailed P value equals 0.0410

The association between rows (groups) and columns (outcomes) is considered to be statistically significant.

Finally, showing percent correct baseline and post in a figure summarizing question results is recommended. For example, for question one, 65.8 percent (25/38) answered correctly at baseline, and 87.1 percent answered correctly at post (27/31). Thus, the absolute increase from baseline to post was 21.3 percent (87.1 percent minus 65.8 percent). However, it is more common to state the relative increase, which would be 32.4 percent, using the formula: [(87.1-65.8)/65.8] x 100. An online calculator for this can be found at Marshu.com.

Working with Paired Data

Having a data set in which the responses to multiple-choice items are assigned to specific individuals is definitely a preferred situation. Such a scenario allows you to consider data from only those learners that offered a response to a question at baseline and at post-activity. Working with a set of data that is restricted in this way, is called working with "paired data." Generally, statisticians think of this as cleaner data that allows for a more powerful analysis to definitively quantify change.

As with unpaired data, the first step is to calculate the group baseline correct percentage and the group post-activity correct percentage to determine the delta for the group being considered. At that point, however, a distinct test is required to calculate the P value. As was shown with unpaired data, the best way to describe the calculations is to show an example.

Case: Paired Baseline/Post Multiple-Choice Question Data

In this case, assume a recent data set for your online educational activity had five outcomes questions that were asked within the delivery of content to assess changes in competence. 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 4** (see page 9).

All changes appear positive, and you shared them with the course director. The course director then indicates a desire to understand the statistical significance of these findings.

Step 1: Access a statistical computation tool.

In order to determine a *P* value for paired data, several tests are available. An easy one to use with free access is found at GraphPad. In order to access the appropriate tests to analyze the data found in Table 4, go directly to the McNemar's test Web page on GraphPad. **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.

In order to put your data into this tool, a bit of data organization is required. There are four possible results when a learner responds to a multiple-choice question twice. First, the learner can answer incorrectly (I) at baseline and correctly (C) at post; for use of this tool, this highly desirable outcome is referred to as "Control = No and Case = Yes." The first cell in the GraphPad tool is for the number of times that this situation occurred. Second, the learner can answer correctly (C) at baseline and incorrectly (I) at post; the number of times that occurs goes in the second cell of the tool corresponding to "Control = Yes and Case = No." Third, the learner can answer correctly (C) at baseline and correctly (C) at post ("Control = Yes and Case = Yes"). The number of times this "reinforcement" finding occurs goes in the third cell in the tool. Finally, a learner can answer incorrectly (I) at baseline and incorrectly (I) at post ("Control = No and Case = No"), and the number of times that occurs goes in the fourth and final field in the tool. A click on "Calculate" returns the *P* value for paired data, as well as several other pieces of information.

To see how this functions, **Table 5** shows the four different scenarios described above for the five questions presented in Table 4. "No/Yes" refers to the count of individual learners who missed the question at baseline but selected correctly at post.

Step 3: Load your data and execute calculation.

It may take a bit of time to prepare your data for the calculation, but once a table like Table 5 is created, plugging the data into GraphPad is fairly simple. An example is shown for Question 1 in **Figure 4** (see page 10).

The key value is the "two-tailed *P* value" determined as 0.6069 for the example Question 1, which is shown framed in a red box in Figure 4. When McNemar's test is performed for all five example questions, we can add *P* values to our original table, shown as in **Table 6** (see page 10)**.**

It is necessary to verify that the *number of discordant pairs is greater than 20* in order for this calculation to be valid. That value can be found in the summary narrative from the GraphPad calculation tool, shown framed in a yellow box in Figure 4. This is an important distinction, as Question 1 has only 34 discordant pairs, even though the response count (n) is 123.

Step 4: 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 123 total matched responses — an "n" that you might expect to give rise to a significant finding. While the delta is only 4 percent, you might be tempted to say that the change is significant and would be greater if there was a lower baseline. However, the *P* value does not confirm that analysis. Conventional criteria would suggest that this positive 4 percent change is fairly random and meaningless.

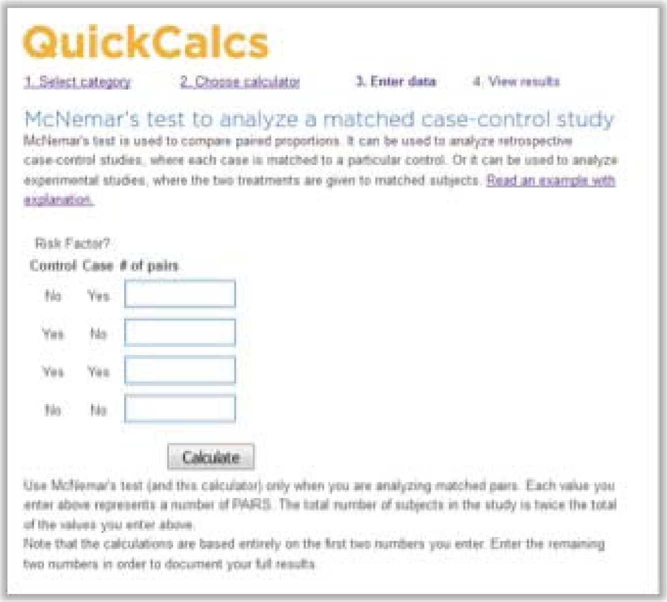
Table 4. A set of paired data from a typical educational activity; n = number of learners responding to both the baseline and post instance of the question

Question #	n	Baseline average correct	Post average correct	Change (D)
1	123	76%	80%	+4%
2	119	51%	77%	+26%
3	51	51%	88%	+37%
4	36	17%	33%	+16%
5	36	53%	94%	+41%

Table 5. Data from Table 4 organized to show the count of learners according to the four different possible ways that multiple-choice items can be answered twice

Question #	n	No/Yes	Yes/No	Yes/Yes	No/No
1	123	19	15	79	10
2	119	38	7	54	20
3	51	21	2	24	4
4	36	7	1	5	23
5	36	15	0	19	2

Figure 3. McNemar's Test Input Screen



QuickCalcs

1. Select category 2. Choose calculator 3. Enter data 4. View results

McNemar's test to analyze a matched case-control study

McNemar's test is used to compare paired proportions. It can be used to analyze retrospective case-control studies, where each case is matched to a particular control. Or it can be used to analyze experimental studies, where the two treatments are given to matched subjects. Read an example with explanation.

Risk Factor?

Control	Case	# of pairs
No	Yes	
Yes	No	
Yes	Yes	
No	No	

Use McNemar's test (and this calculator) only when you are analyzing matched pairs. Each value you enter above represents a number of PAIRS. The total number of subjects in the study is twice the total of the values you enter above.

Note that the calculations are based entirely on the first two numbers you enter. Enter the remaining two numbers in order to document your full results.

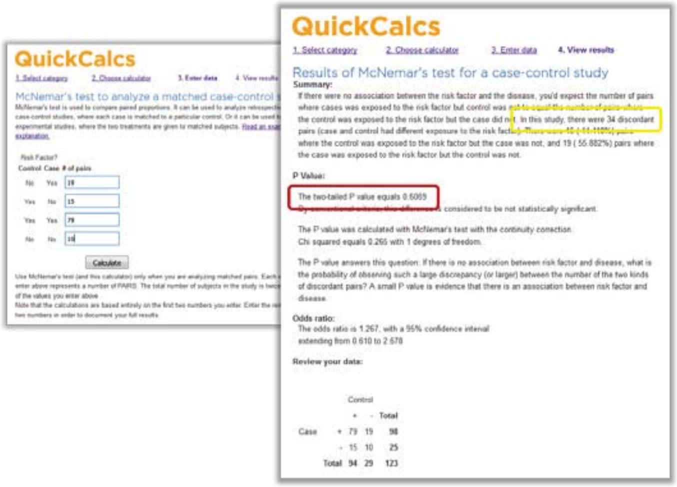
Question 2, on the other hand shows a highly statistically significant finding. Large deltas, when combined with large n's, typically do lead to a statistically relevant finding.

Questions 3, 4 and 5 are also included here for specific reasons. With Question 3, the number of discordant pairs is 23 (n=51), barely allowing for validity of the calculation. What that indicates is that many learners didn't change the way that they responded to this question item, so it's fair to say that many of the 51 percent of learners who got the question correct at baseline had their choice reinforced.

Question 4 has only 8 discordant pairs (n=36) and a *P* value that falls slightly higher than the threshold (< 0.05, as previously mentioned) that most use to qualify for statistical significance. If the activity is ongoing, it may be wise to await additional data. This type of finding is sometimes referred to as a "change trending toward significance."

Question 5 looks very significant with a +41 percent% delta and a *P* value of 0.0003. Unfortunately, the number of discordant pairs is 15 (n=36), which falls short of the needed 20. Because of the *P* value, you might describe this as a "change that is likely to reach significance with additional sampling." As with Question 4, waiting for additional data may address this, if the possibility of additional data collection exists.

Figure 4. A sample calculation in GraphPad with the results page shown at right



QuickCalcs

1. Select category 2. Choose calculator 3. Enter data 4. View results

McNemar's test to analyze a matched case-control

McNemar's test is used to compare paired proportions. It can be used to analyze retrospective case-control studies, where each case is matched to a particular control. Or it can be used experimental studies, where the two treatments are given to matched subjects. Read an exa explanation.

Risk Factor?		
Control	Case	# of pairs
No	Yes	19
Yes	No	15
Yes	Yes	79
No	No	10

Calculate

Use McNemar's test (and this calculator) only when you are analyzing matched pairs. Each enter above represents a number of PAIRS. The total number of subjects in the study is twice of the values you enter above.

Note that the calculations are based entirely on the first two numbers you enter. Enter the re two numbers in order to document your full results.

QuickCalcs

1. Select category 2. Choose calculator 3. Enter data 4. View results

Results of McNemar's test for a case-control study

Summary:

If there were no association between the risk factor and the disease, you'd expect the number of pairs where cases was exposed to the risk factor but control was not to equal the number of pairs where the control was exposed to the risk factor but the case did not. In this study, there were 34 discordant pairs (case and control had different exposure to the risk factor). There were 15 (44.118%) pairs where the control was exposed to the risk factor but the case was not, and 19 (55.882%) pairs where the case was exposed to the risk factor but the control was not.

P Value:

The two-tailed P value equals 0.6069

By conventional criteria, this difference is considered to be not statistically significant.

The P value was calculated with McNemar's test with the continuity correction.
Chi squared equals 0.265 with 1 degrees of freedom.

The P value answers this question: If there is no association between risk factor and disease, what is the probability of observing such a large discrepancy (or larger) between the number of the two kinds of discordant pairs? A small P value is evidence that there is an association between risk factor and disease.

Odds ratio:

The odds ratio is 1.267, with a 95% confidence interval
extending from 0.610 to 2.678

Review your data:

		Control		
		+	-	Total
Case	+	79	19	98
	-	15	10	25
	Total	94	29	123

Table 6. Addition of calculated *P* values to assess significance of change percentages (Δ)

Question #	n	Baseline average correct	Post average correct	Change (Δ)	*P* Value
1	123	76%	80%	+4%	.6069
2	119	51%	77%	+26%	<.0001
3	51	51%	88%	+37%	.0002
4	36	17%	33%	+16%	.0771
5	36	53%	94%	+41%	.0003

Limitations

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 responses 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.

For paired data, we mentioned several times that McNemar's test has a minimum number of discordant pairs limitation. There are other calculations that avoid this specific limitation, but for the purposes of this beginner's data analysis article, we've chosen to propose the use of McNemar's test as it covers most cases and generally works quite well when paired n's are higher than at least 40 on an individual question.

Summary

The *P* value represents the role that chance plays in your outcomes. Researchers accept that chance may play some role in their findings, but only if that chance is 5/100 (*P* = 0.05) or less. Any greater likelihood of something happening due to chance is grounds for saying that your findings are random.

With the right tools, setting up and calculating the statistical significance of your findings is really fairly simple, and you can make it easily repeatable if you have the inclination to better understand what your data have to say. One topic that hasn't been addressed elsewhere in this article is the issue of very low n's; what if you have very few respondents (e.g., 15 at baseline and 10 at post)? Low participation in analysis of multiple-choice outcomes data does present a challenge that is not easy to overcome.

In short, the lower the number of your respondents, the larger the change from baseline to post in correct answers needed to achieve statistical significance. In the absence of enough data (typically no fewer than 30 responses at baseline and/or post is needed to give yourself a chance at measuring significance using a credible statistical test like Chi-square), our recommendation is to show relative increase in correct answers from baseline to post.

References

1. http://www.dummies.com/how-to/content/statistical-significance-and-pvalues.html Accessed 10/23/15.

2. http://www.dummies.com/how-to/content/what-a-pvalue-tells-you-about-statistical-data.html Accessed 10/23/15.

Resources

1. Jason Olivieri, CMEPalooza, Statistical Analysis in CME Outcomes, http://cmepalooza.com/march21/statistical-analysis-in-cme-outcomes-olivieri/

2. 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=11l75UrlqxE