



OMNI

GUIDE TO INTERPRETING

(AND MAKING THE MOST OF)

YOUR INCREDIBLE YEARS DATA

OMNI's philosophy is that evaluation results are most useful when they are *understood and utilized* in order to improve service delivery. We analyzed your data using paired-sample t-tests to measure the level of statistical significance and effect size in observed differences between the pre-test and post-test means. Although all necessary statistical results are provided within provider-level data tables, individualized interpretation is not provided. Therefore, we have provided this guide as a means for you to:

- 1) Learn how to review your individual data table results and
- 2) Use knowledge gained towards improvements in services or practices

It is understood that no manual or guide can effectively “teach” someone how to interpret complex statistical data results. However, it is hoped that this guide serves as a fundamental introduction to key terms, concepts, and questions to consider when exploring your data results.

Here are definitions for terms that will be helpful for you in reviewing your results:

Items: The individual questions in a scale.

Scale: The collection of related items (questions) that measure a specific construct (i.e. attribute, character—for example, Parent/Child Bonding).

Measure: Term typically used interchangeably with Scale.

Sample: The respondents included in your data set.

Overall Mean: The mean score of all the items (questions) in the scale.

Item mean: The mean score of one item (question) in a scale.

Where applicable, it is more meaningful to report the scale mean because it takes into account all of the items (questions) that measure the main construct.

“R”: When an item or survey question is marked with an “R” it means “reverse-scored”. To create overall means, some items may need to be reverse-scored so that all items have the same expected direction of change. Usually, but not always, negatively worded items are reversed so that in an overall mean, “more is better”.

Pre: Reference to the pre-test data

Post: Reference to the post-test data

Change: The difference between pre-test and post-test means

P-value: The probability that a difference between two means is due to chance alone.

Program Effects: What you would like the data to reflect as a result of your program (for example, an increase in perception of risk and harm of substance use.)

Effect Size: How much of an effect did the program have?

A paired samples t-test analysis examines the difference between the pre- and post-test means to determine whether an observed difference is statistically meaningful. The analysis calculates the probability (p-value) that the changes in the scores of the pre-test and post-test are due to chance rather than being a real change.

Standard practice in the social sciences is to consider p-values less than (<) 0.05 statistically significant. This basically says that social scientists will only conclude that a difference between two means is meaningful when there is less than a 5% probability that the difference is due to chance

alone. In some cases, especially when sample sizes are small, p-values between .05 and .10 are worth noting because they approach the .05 benchmark. In these cases, the term “approaching significance” will be used.

A limitation of interpreting the p-value and statistical significance is with regard to statistical power. **Most small programs lack an adequate sample size** to evoke confidence in the p-value and **test of statistical significance**. A more unrestricted analysis is to determine the effect size to answer the question: how much of an effect did the program have? Effect size analyses provide an indication of the amount of change regardless of sample size. Effect size can be interpreted similarly to a “percent difference”, on a metric between .00 and .99. Effect sizes can be negative or positive, and a score of 0 represents no change. Generally speaking, effect sizes in social research are likely to be small (under .2).

The following is a set of benchmarks to determine the magnitude of effect sizes for your program:

.20	=	small effect size
.50	=	moderate effect size
.80	=	large effect size

Effect sizes and p-values can be used together to provide a more comprehensive picture of true program outcomes. **In the case of a sample size of 15 or fewer, p-values should not be interpreted.** In those instances, effect sizes can provide a “benchmark” for comparison against other small sample results. Effect size is another standard practice used in the social sciences.

What does it mean if my results are statistically significant (or if the mean differences have large effect sizes)?

Statistical significance and/or moderate-to-large effect sizes might mean that your program is doing what you hoped it would. However, it is important to consider whether other factors, unrelated to your program, might have had an impact on the results. For example, if the change in bonding (as measured by the mentoring pre-test and post-test) was desirable, it is possible that your services were part of the reason for the change. However, it is also possible that these youth experienced something else during this time period that also had an impact on their bonding with school and community. For example, perhaps their school just received additional funding for new facilities and teachers and this influx of financial support translated into more interesting and exciting programs at the school. While you want to celebrate the success, it is always important to interpret statistical results with caution. Most importantly, continually think about ways to improve your evaluation-related practices as well as program services.

What does it mean if my results are NOT statistically significant (or if my effect sizes were very small)?

Remember, guidance on **statistical significance applies only to those agencies that have sample sizes of larger than 15**. If your agency has fewer than 15 matched participants, you are unlikely to demonstrate statistical significance *due to small sample size*.

Just as it is important to use caution when attributing statistical significance or large effect sizes to your program, it is important to understand that, sometimes, positive effects are difficult to detect. If some evaluation results are *not* statistically significant or have a small effect size, this does not necessarily mean that your program did not cause change in your participants.

Just a few reasons may be:

Evaluation Implementation

- Your sample size (the number of participants who took your pre-test and post-test) was too small for a change to be detected. Did you accurately track participants from pre to post? Did too many participants drop out (attrition) from your services? Did you lose data due to data-entry error?
- Participants weren't completely honest in their responses. Did you ensure confidentiality? Did you make sure participant names were not included on their surveys? (But, make sure that you have a tracking system in place.)
- Your measure was not the best fit to capture the change. Is it reliable and valid? Does it measure outcomes directly tied to your services? Does it provide a range of response options?

Program/Population

- Your program may not be sufficient to make a significant effect on the participants. Do you want to consider adding or supplementing services? Do you have the right combination of services? Were you implementing the program with fidelity? That is, did you deliver enough of the program in the way it was intended?
- You did not target the population you intended. For example, you are measuring change in attitudes towards use. As it turns out, youth in your program already come in with very high negative attitudes towards use – meaning that there is no room for even more negative attitudes. Thus, no statistically significant results from pre to post test. (This is often called a “ceiling effect.”) However, you thought you had targeted higher-risk youth.

How to use this guide

In order to best use your pre-post test data, you will want to go through your data table(s) carefully; asking and answering questions along the way such as “Did I have a large enough sample size and if not, why not?, Is the change what I thought I would see?”, and many others. Answers to these questions may guide several action steps for you to take regarding your Programs and Services.

Here is an example of your site-level data table you will receive:

[Note: there is a blank worksheet to follow for interpretation of your program’s data below this example for your use.]

Number of Matched Pre-Post Tests: 25

The questions below were answered according to the following scale:

- 1 = not at all
- 2 = a little
- 3 = moderately well
- 4 = well
- 5 = very well

Program Effects: You’d like to see an increase in the mean from pre- to post-test, indicating that teachers report an *increase* in social competence among their students.

SCST Scales	n	Pretest Group Mean	Posttest Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Prosocial/Communication Skills (PCS) Scale	25	3.04	3.77	0.73	yes	*	0.80
Emotion Regulation Skills (ERS) Scale	24	3.52	3.85	0.33	yes	*	0.35
Academic Skills (AS) Scale	24	3.02	3.15	0.13	yes	*	0.14
PCS + ERS Combined Scale	25	3.21	3.99	0.78	yes	*	0.86
Overall Scale Mean	25	3.10	3.75	0.65	yes	*	0.73

n.s. indicates the difference is not statistically significant

^ indicates the difference is approaching statistical significance (.05 < p < .10)

** indicates the difference is statistically significant (p < .05).*

The next few pages walk you through how to read the data table:

What is the effect size? .73

Look at the last row of the data table. That column will tell you the effect size.

SCST Scales	Pre-test Group Mean	Post-test Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Overall Mean	3.10	3.75	.65	Yes	*	.73

Are your results statistically significant? Yes

Look at the second to last row of the data table. The column marked “significance” shows whether your results are statistically significant or not.

SCST Scales	Pre-test Group Mean	Post-test Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Overall Mean	3.10	3.75	.65	Yes	*	.73

What is the pre-test mean? 3.10

SCST Scales	Pre-test Group Mean	Post-test Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Overall Mean	3.10	3.75	.65	Yes	*	.73

What is the post-test mean? 3.75

SCST Scales	Pre-test Group Mean	Post-test Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Overall Mean	3.10	3.75	.65	Yes	*	.73

What is the Mean Change from pre to post? .65

SCST Scales	Pre-test Group Mean	Post-test Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Overall Mean	3.10	3.75	.65	Yes	*	.73

Is this the desired direction of change? Yes

*Remember, even if your data is not statistically significant, you can still see the desired direction of effects. Conversely, statistically significant results can be in the **wrong direction** (e.g., teachers reported a decrease in social competence).*

SCST Scales	Pre-test Group Mean	Post-test Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Overall Mean	3.10	3.75	.65	Yes	*	.73

What does a mean of 3.10 indicate on this scale?

(e.g., “It’s between a 3 and a 4, but closer to a 3. So, a 3.10 indicates that teachers report moderately well social competence among their students”)

<p>SCST Scales</p> <p>1 = not at all</p> <p>2 = a little</p> <p>3 = moderately well</p> <p>4 = well</p> <p>5 = very well</p>

“Making the Most of Your Pre-Post Test Data” Program Data Worksheet

Take a close look at the overall and item means...

What is the pre-test mean? _____

What is the post-test mean? _____

Look at your scale. Write out what changes you'd like to see.

(e.g., “I'd like teachers to see an increase in social competence. I'd like the mean to go up to a 4 or 4.5.”)

What is the Mean Change from pre to post? _____

Did you see the desired direction of change? _____

(Remember, even if your data is not statistically significant, you can still see the desired direction of change. Conversely, statistically significant results can be in the *wrong direction!*)

Is the change higher or lower than you expected? _____

Now look more closely at your scale...

Revisit your pre-test mean. What kind of room is there to reflect change from pre-test to post-test?

(Were the pre-test scores low enough to allow room for improvement? That is, if your participants reported a mean of 4.75 at the beginning of the year and the scale only goes to 5, that doesn't leave much room to improve.)

If not enough room for change on the scale, is that ok? Were you hoping to maintain a certain level, as opposed to increasing or decreasing scores?

(e. g., It might be that your program could seek to keep kids at the same level throughout the year, in which case – success!)

Now look at significance and effect size...

Are your results statistically significant? _____

How large is the effect size? _____

Is the effect size small, moderate, or large? _____
(.19 and under =small .20 -.49=moderate .50 or above = large)

Taken together, what do both the statistical significance test **as well as** effect size tell you about the possible impact of your program?

Almost Done! Now apply lessons learned from the data to thinking about programs & services...

What do the outcome results tell you? (i.e., briefly summarize your results)

Overall, were the pre-post test differences what you expected? And if not, how were they different?

If the results were not what you expected, what factors should be considered?
(e.g., lost data due to evaluation protocol problems, did not provide the full dosage of the program, etc.)

Look at your pre-test mean. Does it reflect who you originally targeted?
(e.g., Do the pre-test scores surprise you? Is the population served different than your targeted population?)

What thoughts arise regarding program implementation based on what you learned?
(e.g., if your participants are higher-risk than you anticipated, are there any implications for program implementation? Would you choose to supplement services?)

Are there any evaluation practices you might change based upon lessons learned?
(e.g., Put better client ID tracking systems in place? Consider a better measure?)

Comparison Means – what do they mean?

This section only applies to you if you have been provided with comparison results. If not, feel free to skip this section now and refer back to it later if comparison data become relevant.

Utilizing comparison means allows a program to not only look at what levels of change you expected from your clients but can put your results and information into a larger context. This does not mean the comparison results are nationally representative, or the only “correct” results. You know best the special features of the population you serve. Consider each item individually when you think about whether you would expect individuals from your organization to be higher, lower, or about the same as results from a larger and more diverse population.

Statistical significance is strongly influenced by sample size and comparison results will, by definition, have a larger sample size than those from your agency. Pay more attention to the **mean difference**, and to the **effect size**. If the mean difference from your agency is in the desired direction, and the same size or larger than the comparison mean difference -- the *smaller sample size is the only reason why those results were not statistically significant*.

Below you will find an example worksheet with program-level data (from page 5) and overall data already completed, which will walk you through this process.

Comparison-Aggregate data

Social Competence Scale Teacher Aggregate Data

Number of Matched Pre-Post Tests: 949

The questions below were answered according to the following scale:

1=not at all, 2=a little, 3=moderately well, 4=well, 5=very well

SCST Scales	n	Pretest Group Mean	Posttest Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Prosocial/Communication Skills (PCS) Scale	948	3.07	3.65	0.58	yes	*	0.63
Emotion Regulation Skills (ERS) Scale	948	3.04	3.61	0.57	yes	*	0.61
Academic Skills (AS) Scale	949	3.19	3.75	0.56	yes	*	0.60
PCS + ERS Combined Scale	949	3.05	3.63	0.58	yes	*	0.64
Overall Scale Mean	949	3.09	3.68	0.59	yes	*	0.66

* indicates the difference is statistically significant ($p < .05$).

^ indicates the difference is approaching statistical significance ($p > .05, < .10$)

n.s. indicates the difference is not statistically significant

Here are questions you will want to work through when looking at your data in comparison to the statewide data:

Now that you have reviewed and summarized your individual program-level data, look at the comparison data. What are the differences that you see right away?

(In this example, at first glance, you see that the aggregate data showed a desired direction of change on all items and so did your program data.)

Compare your program’s pre mean scores to the comparison pre mean scores. What are the differences?

(In this example, there are no notable differences in pre-test group means between your program-level data and the comparison-aggregate data.)

Compare your effect size with the effect size of the comparison-aggregate data results.

Remember that effect size analyses provide an indication of the amount of change regardless of sample size. Therefore, comparing your effect size with the aggregate effect size can give an indication of how your agency is doing relative to all agencies combined.

(For example, the effect size for your program was .35, which suggested a moderate positive impact on youth’s intentions not to use substances. The effect size for the aggregate data was .66, which is a large effect size. This indicates that across all agencies the program had a large positive impact on the youth, which is a bigger impact than your program.)

What are you able to conclude from the comparison means that you might not have seen in your data before?

Program data:

Emotional Regulation Skills (EMR) Scale	3.52	3.85	.33	No	*	.35
---	------	------	-----	----	---	-----

Overall data:

Emotional Regulation Skills (EMR) Scale	3.09	3.68	.59	Yes	*	.66
---	------	------	-----	-----	---	-----

Below are the aggregate results for the Parent Practices Interview. The above exercise can be conducted using this data to compare to provider-level results.

Comparison-Aggregate data

Parenting Practices Interview Aggregate Data

Program Effects: You'd like to see an *increase* in the mean from pre- to post-test, indicating that parents report using more positive parenting techniques with their children.

The questions below were answered according to a seven-point scale.

Positive Parenting Scales	n (pre,post)	Pretest Group Mean	Posttest Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Appropriate Discipline (AD)	183,129	4.26	4.80	0.54	yes	*	0.57
Positive Parenting (PP)	183,129	4.53	5.00	0.47	yes	*	0.64
Clear Expectations (CE)	177,124	5.28	5.74	0.46	yes	*	0.43
Monitoring (MO)	183,129	5.93	6.07	0.14	yes	^	0.08

n.s. indicates the difference is not statistically significant

^ indicates the difference is approaching statistical significance ($.05 < p < .10$)

* indicates the difference is statistically significant ($p < .05$).

Program Effects: You'd like to see a *decrease* in the mean from pre- to post-test, indicating that parents report using less negative parenting techniques with their children.

The questions below were answered according to a seven-point scale.

Negative Parenting Scales	n (pre,post)	Pretest Group Mean	Posttest Group Mean	Mean Change	Desired Direction of Change?	Significance	Effect Size
Harsh Discipline (HD)	183,129	2.46	2.00	-0.46	yes	*	0.65
Harsh for Age (HFA)	182,128	1.69	1.70	0.01	no	n.s.	-0.01
Inconsistent Discipline (ID)	183,129	2.98	2.51	-0.47	yes	*	0.19

n.s. indicates the difference is not statistically significant

^ indicates the difference is approaching statistical significance ($.05 < p < .10$)

* indicates the difference is statistically significant ($p < .05$).