In the BU354 Comparative Usability Test, a paired samples t-test is performed to compare the rankings of two dashboards. The Python code the uses `scipy.stats.ttest_rel` package. Internally it computes the difference of the observed difference between the scores of the two dashboards.

$$d_i = x_i - y_i$$

The mean of these paired differences is then calculated as:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$$

Since the data represent a sample of some students as opposed to the population of students, the sample standard deviation of the differences is used:

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$

The paired t-statistic computed by `ttest_rel` is equal to to:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \qquad df = n - 1$$

This difference compares the observed mean difference to the standard error of the mean difference. A small standard error (typically produced by larger sample sizes or low variability) leads to a larger magnitude of the t-value.

In addition to statistical significance, we also compute effect size using Cohen's d for paired samples, which uses the standard deviation of the difference scores:

$$d_{\text{Cohen}} = \frac{\bar{d}}{s_d}$$

This value indicates the magnitude of the difference independently of sample size. The Python code mirrors this formula directly by dividing the mean difference by the standard deviation of the paired differences.