STATISTICAL SIGNIFICANCE AND THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Why does NAEP reporting include references to statistical significance?

NAEP provides us with performance results for large groups of students without testing every student. Instead of testing every student, NAEP uses a complex sampling design to select representative groups of students for testing. So, in Connecticut not every school is selected for NAEP and within the schools selected for NAEP, it would be unusual to test every student. Even though NAEP tests a sample of Connecticut students, the program is able to report results for the state and subgroups of students within the state. This process of sampling schools and students reduces the burden on schools and increases the efficiency of the test administration overall.

By testing representative samples of students, NAEP is able to provide performance estimates for the nation, states, and subgroups. However, it is important to understand that whenever we select a sample and report results for a population, there will be variability from the total population value depending on the sample selected. This variability is referred to as statistical error. For example, political polling is designed to determine what a population thinks about an issue or candidate. The polling is conducted with a sample of the population and results typically are reported along with a *margin of error*. NAEP also uses a *margin of error* but presents the information in a slightly different manner. Rather than provide an interval or range of performance (e.g., the average scale score is 280 plus or minus five points), NAEP reports *standard error* values with all results.

In NAEP, the standard error values help people determine the amount of variability in the results that are reported. NAEP goes one step further in clarifying the information for the public by reporting all results in terms of statistical significance. Therefore, when NAEP states that one group of students is achieving proficiency at a higher rate than another group, the reader can be confident that there is a statistically significant difference (i.e., the reported results exceed the margin of error). NAEP does not make statements claiming performance differences unless there is a statistically significant difference. In other words, NAEP will not highlight an apparent difference unless the difference exceeds what we would expect due to variation (or error) that is a result of testing a sample of students rather than the entire population. This means that two states could have different average scale score values (e.g., 275 and 277), but there may not be a statistically significant difference because of the standard errors. As a result, NAEP will not claim that the average scale score of 277 is higher than the 275 value. Instead, the reporting will indicate that the states are performing at the same level or that the results are not statistically different.

All NAEP reports issued by the Connecticut State Department of Education (CSDE) follow the same reporting conventions as the official NAEP reports issued by the National Center for Education Statistics. The CSDE will not claim changes in performance unless there are statistically significant differences.

When the nation's average scale score increases by one point, the results indicate that the nation improved. However, when Connecticut's average scale score increases by one point, the results indicate that Connecticut's student performance stayed the same. Why?

Getting back to the political polling example, when the pollsters sample large groups of people, the margin of error is very small and when they sample small groups of people, the margin of error can be quite large. The same is true in NAEP. The standard error values for the national results are very small because the sample at the national level is extremely large. In contrast, a subgroup of students in Connecticut would likely have relatively fewer students in the state sample. Therefore, the standard error values for small subgroups of students in Connecticut would be quite large, which means that there is a large margin of error around the results for small subgroups.

Large standard error values sometimes make interpretation of change over time difficult to understand. When comparing performance in 2017 to performance in 2015, small subgroups of students have to show quite a large change in performance to overcome the large standard errors and show a statistically significant difference. Whereas, large groups of students can show a small average scale score point gain and the results are interpreted as statistically significant.

Are all statistically significant changes important changes?

Not necessarily. Statistical significance helps us to determine whether two values are different. The magnitude and importance of a difference is a matter of context and interpretation.