RESEARCH BRIEF
Research Services

Vol. 0806
April 2009

Dr. Terry Froman, Research Services
Dr. Aleksandr Shneyderman, Office of Program Evaluation

MEASURING FCAT IMPROVEMENT

Since the inception of the FCAT, the scaled scores students receive have been summarized by categorizing them into levels of performance. For every single test (in one subject area, at one grade level, for one year) the State of Florida has carefully determined cutoff scores that classify performance into 5 different levels of proficiency. By now, we are familiar and comfortable with characterizing test performance in terms of performance levels. It seems quite natural to refer to the percentage of students beyond a particular cutoff, as when we say a certain percentage of students scored at Level 3 or above. However, this seemingly straightforward "percent-above-cutoff" type of summary statistic when used to describe improvement in performance has subtle built-in difficulties that can lead us to grossly inaccurate inferences.

There are alternative ways of summarizing FCAT performance that are just as simple and easy to interpret as percent-above-cutoff type descriptions but do not carry with them the inherent dangers. It is the purpose of this paper to show exactly when and how our commonly used FCAT reports can go wrong and suggest alternative statistics that keep us on the right track.

Reporting Results

We can refer to the test performance of a group in two different ways.

- The **percent description**: referring to the percent of students scoring above some score cutoff (like "62% of the students scored above a score of 322" or "71% of the students are at Level 3 or above"), and
- The **score description**: referring to the score that corresponds to some particular percentile point (like "the median percentile was a score of 307" or "the 90th percentile was represented by a score of 412").

On the surface, these look like logically equivalent ways of reporting group test performance. Whether you start off choosing the score of 322 and find the corresponding percentile of 62% or you choose the 62nd percentile and find the corresponding score of 322, you end up with the same description: "The score of 322 goes with the percentile of 62."

But, the distinction between these two approaches to summarizing scores is more than just a matter of semantics. When you start talking about **differences** in group test performance, the way you report summary data can matter. This is true whether you are talking about comparisons between two different groups at the same point in time or between two different times for the same group.

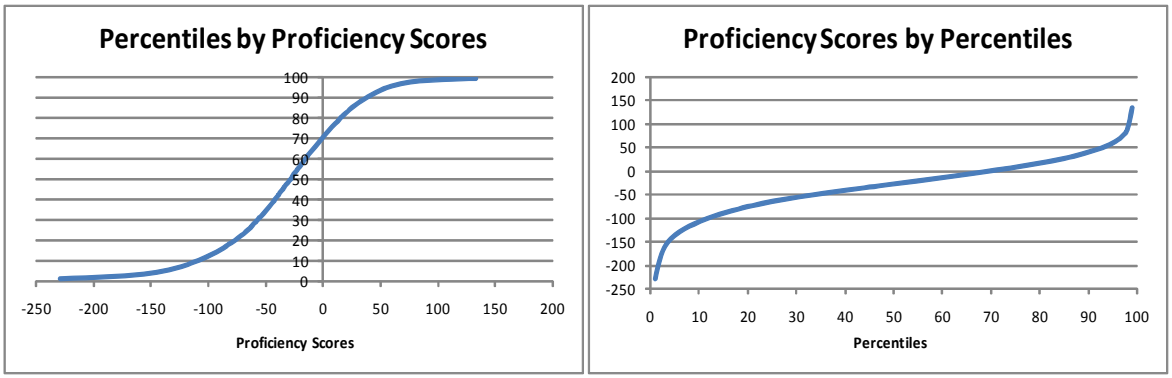
Here is one way that the problem can surface. Imagine that we set a goal that "schools should improve by 5 percentage points in the percent of students scoring at Level 3 or above." In some school that starts out with **45** percent of their students at Level 3 or above, their mean scaled score might have to improve by **5** points to achieve that goal. At another school that starts out with **75** percent of their students at Level 3 or above, their mean scaled score might have to improve by an average of **12** points to achieve the goal. The same goal definition, in terms of percent gain, can be very different challenges in terms of score improvement for schools starting at different places.

On the other hand, had we stated the goal in terms of the score at a percentage point (i.e., "schools should improve by 5 scaled scores at their 50th percentile point") things would be different. Now, the change in score value (whether we reference the 50th percentile or some other percentile point) would be the same and represent the same academic challenge.

A Graphical Illustration of the Two Approaches

The two graphs below depict the orientation of each of the two approaches. On the left is the cumulative distribution of the scores for a single group of students, Grade 5 at the District level. The difference between the scaled scores and the proficiency cutoff is represented on the horizontal axis and is denoted as the "proficiency score." One can track up from any particular proficiency score to its intersection on the graph curve, and from that point to the vertical axis to locate the percent of students scoring up to that score point. A typical score point of interest would be the percent scoring at Level 3 or Above, which corresponds to the difference of zero to the proficiency cutoff score. It's easy to see that the corresponding percent of students scoring up to this cutoff for this example is approximately 70 percent.

In the graph on the right is the same cumulative distribution, but here oriented with the percentiles on the horizontal axis and the difference between the scaled scores and the proficiency cutoff on the vertical axis. From this perspective one may be interested in the score that corresponds to a chosen percentage point. A common choice here in the graph is the score which corresponds to the 50th



percentile: approximately a score of -25, or twenty-five scaled score points below the proficiency cutoff.

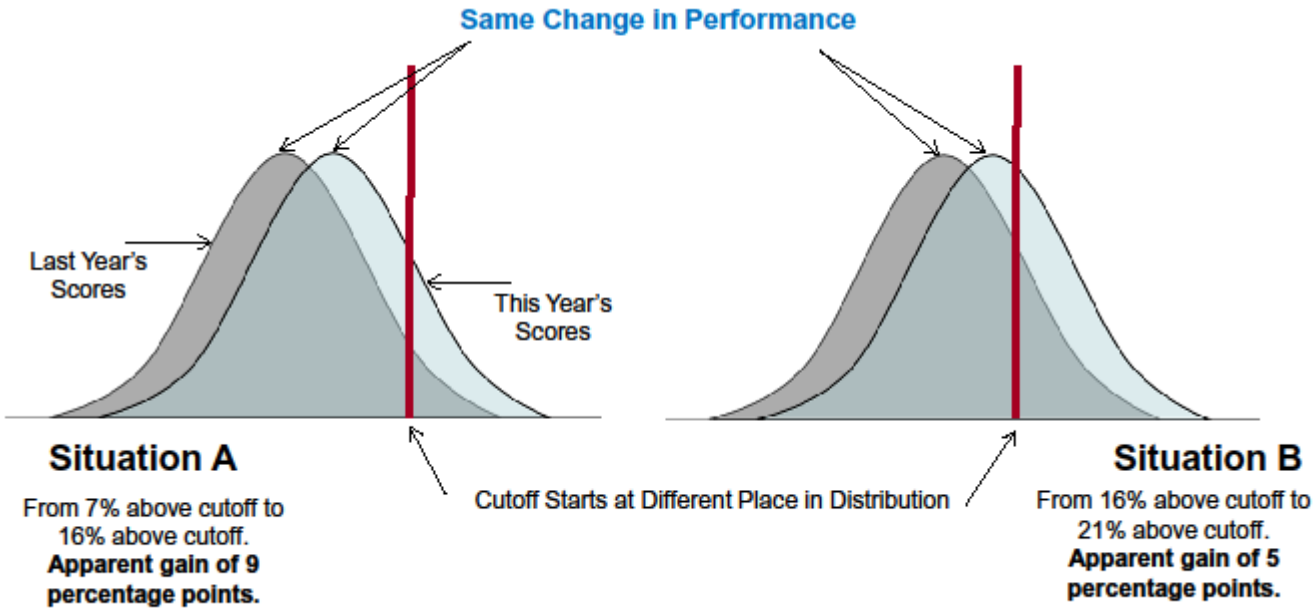
We are trying to emphasize a difference in referencing scores versus referencing percentages, but the graphs are really very similar – the axes are interchanged. Any information you could glean from one graph, you could easily derive from the other. The score below which 50 percent of the students score is easily found in the left graph, and the proficiency difference of zero in the right graph corresponds to an easily identifiable percent of students in the graph to the right. To advocate one

approach over the other when exploring the results for a single group of students wouldn't make much sense. But, when we start to compare results between different subgroups or across time, the choice of an approach matters.

The Reason for the Problem with Percent-Above-Cutoff Thinking

There is a technical difference in interpretation between changes in score points and changes in percentages that is not readily apparent but, nonetheless, important. For any particular group, a difference in average points scored on the FCAT test represents the same change in academic performance no matter where it occurs in the scale (that is, a change from an average score of 210 to 220 represents the same increase for the group in academic achievement as a change from 310 to 320). On the other hand, a difference in percent-above-cutoff may represent very different changes in academic performance depending upon where it occurs in the distribution (that is, a change from 50% to 55% above a cutoff is less of an increase for the group in academic achievement than a change from 85% to 90%).

The basic problem, in the simplest of statistical terms, is that the **percent description <u>is not</u> on an equal-interval scale of measurement.** So, differences in percent at particular scores are not comparable at different places in the distribution. However, the **score description <u>is</u> on an equal-interval scale**. So, differences in scores at different places in the distribution are comparable. Although this might not be true for all tests, this is how these test scales were originally designed.



The apparent gain in percentage points is partly a function of the location of the cutoff point in the distribution, as the following figure illustrates.

Imagine an idealized state of affairs where each student this year has scored exactly 15 scale points higher than they did last year. If we try to describe their improvement in terms of percent above a cutoff point, the interpretation can change depending on the placement of the cutoff point. Situations A and B, graphed above, show the consequences of the different cutoff points.

In both situations depicted above, the distributions have shifted **the same amount** to the right between last year's and this year's scores, indicating the same amount of gain in score points for every student. That this represents equal gain in the underlying trait being measured (in this case, academic achievement) is especially true for test scores based on Item Response Theory, as are the FCAT scores. But when the cutoff scores are in different locations relative to the distributions, as they are in the two situations depicted above, the amount of increase as measured by percent-above-cutoff calculations appears to be different for the two situations. Clearly, the percent-above-cutoff description is providing misleading information.

A description of these same differences of students' scores in score-at-percentage-point terms would reveal the true nature of the changes. No matter what percentage point is chosen for reference, the difference in corresponding score from last year to this year would always be exactly the same – the 15 points each student has gained.

How Should We Report Performance?

As implied earlier, if you're just looking at one group at one point in time, either method of reporting is fine. But inevitably, we will be reporting results for different groups and different times. And, we will inevitably want to make comparisons. For these comparisons to be trustworthy, we need to begin thinking of **scores** at certain **percentage points** in the distributions.

Because the situation can be subtly different at various points in the distribution, a complete picture of the group performance is best demonstrated by a complete graph of the distribution. This complete picture approach is exemplified by the kinds of score-at-percentage-point graph presented earlier in this paper. This can be manageable when dealing with just one or two comparisons, either between times or groups. But when we look at trends over several individual comparisons, the graphs can become crowded. We always run the risk of losing information when we summarize, but some type of further summarization is necessary.

We can probably get a realistic picture of the performance trends by checking in on just a few reference percentage points. A nice set might be the 10th, 25th, 50th, 75th, and 90th percentile points. Tracking score changes at these percentile points across advancing grade levels should give us a good sense of development trends. The table below presents the scores at this set of percentiles for grade 3 through 10. Because the proficiency cutoffs vary among grade levels, the scores in the body of the table are the distance each corresponding scaled score is above or below the proficiency cutoff. Thus, negative scores are below minimal proficiency standards and positive scores are above proficiency standards.

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
10th percentile	-61.5	-55.5	-61.5	-71.5	-68.5	-69.5	-91.5	-118.5
25th percentile	-15.5	-15.5	-18.5	-27.5	-25.5	-33.5	-51.5	-74.5
50th percentile	24.5	19.5	17.5	8.5	11.5	-1.5	-15.5	-31.5
75th percentile	59.5	52.5	52.5	43.5	43.5	26.5	15.5	5.5
90th percentile	91.5	80.5	82.5	74.5	74.5	52.5	40.5	37.5

Reading FCAT Trends across Grade Levels

In this table it is easy to see that the performance as measured by the distance from the cutoff generally get worse as the grade levels increase. For instance, the performance score differences associated with the 50th percentile start in positive ground in grade 3, indicating that students in the middle of the distribution perform above the cutoff for Level 3. The proficiency scores associated with the 50th percentile for grades 8, 9, and 10 are progressively more negative, indicating the middle of the distributions are lagging progressively farther behind proficiency standards.

It may be necessary to make a greater level of summarization of the data. If a single reference point must be chosen to represent the performance changes, the familiar 50th percentile would be a logical choice. If the assumption of normality is justified, as is presumed in most testing situations, the commonly available mean scaled score would provide the same kind of information. The important point is, by concentrating on scores at identifiable reference locations in the distribution, our interpretations of improvement trends are accurate and reliable.

Conclusions

We have tried to present several perspectives that illustrate the potential problems with reporting FCAT scores in the customary method adopted by the State of percent-above-cutoff summarizations. The essential difficulty is that **differences** in percents relative to chosen score cutoffs are not trustworthy measures of change. Instead, if we wish to make valid inferences of performance comparisons, we should be reporting score values that correspond to chosen percentile points in the distribution.

The State went through a careful process to choose the cutoff points that define the performance levels and the standards for minimal proficiency. Although there may still be a degree of arbitrariness to the exact cutoff points, this is not the source of the problems we document here with cutoffs. **Any cutoff point**, even the perfect cutoff point, can lead to distorted pictures of performance comparison, change, and improvement.

However, this does not mean that the State's method of proficiency level cutoffs is not useful. It is, after all, quite natural to concern ourselves with the percent of students, as members of any defined group, that have attained a performance level of proficiency. Every school district in Florida has followed the State's model of reporting scores in terms of percents at performance levels and many of our practical concerns have been stated in terms of inferences from this style of reporting. It would be foolhardy of us to suggest that we abandon performance level data. We will still see summary results in terms of cutoffs and set goals in terms of percent proficient. But, now we should be more wary of our interpretations of change and back up those inferences with studies from the more reliable score-at-percentage-point perspective.