RESEARCH BRIEF
Research Services

Vol. 1001
July 2010

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

Single-Scorer School Grading Formula

How to calculate school grades when score of 3.5 is no longer possible in the FCAT Writing results.

Due to budgetary concerns, the FDOE restructured the Writing portion of the FCAT assessment program for 2010. Among the changes was a difference in how the writing essays were scored. In the past, two people independently judged each essay and each assigned a score from 1 to 6. In the event of a disagreement between judges, the average of the scores was assigned. Thus, "half-value" final scores like 2.5 and 3.5 were possible results. When it came to grading schools, the convention was to have the percent of students scoring 3.5 or higher constitute the writing component of the overall point total. However, in 2010 only a single judge would score each essay. Scores between integer values (2.5, 3.5, etc.) would no longer be possible. Since a score of 3.5 could not occur, it would not be possible to summarize a school's performance in an equivalent manner comparable to the "percent scoring 3.5 or higher" standard of the past. To keep the school summary scores as alike in meaning as possible, some accommodation to the school grading methodology would have to be introduced.

Since a numerical school summary defined by the "percent scoring 3.5 or higher" would not be strictly possible, the initial solution considered was to simply substitute the average of the "percent scoring 4 or higher" and the "percent scoring 3 or higher." On the surface, this seems to make sense; after all, the average of 3 and 4 is 3.5. However, while this proposal superficially seems straightforward and fair, there is a subtle source of systematic bias hidden in that kind of computation. The purpose of this paper is to explain the nature of the calculation error and suggest an alternative procedure that would provide a more accurate estimate consistent with the "3.5 or higher" traditional approach.

Example Data

It will be easiest to illustrate the mathematical details of the different grading formulas through an example. For this purpose, we can refer to the actual scores from the 2009 administration of the FCAT Writing test. For the 4th grade Combined scores across the State as a whole the data are as follows.

Number and Percent of Students Earning Each Score Point on the Prompt

	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	
n =	1221	733	3392	4357	19260	23390	79583	33086	20058	6091	2165	193336
percent	0.6%	0.4%	1.8%	2.3%	10.0%	12.1%	41.2%	17.1%	10.4%	3.2%	1.1%	100%

With all of the half-value scores represented, it is easy to calculate the percent scoring 3.5 or higher as 164373 / 193336 = 85.0%.

Converting to Integer-Only Scores

In this set of data from last year, all of the half-value scores were possible. Had these same essays been scored by a single judge, only the integer value scores would have been represented. It is difficult to say exactly what would have happened to the in-between scores. How many of the 3.5's would have reverted to 3, and how many would have reverted to 4? Since the essays would have randomly been assigned to one of many possible judges, there is good reason to suspect that, on average, half of the in-between scores would convert to the lower integer value score and half would convert to the higher integer value score. Thus, an integer-only simulated set of data based on the actual scores would look like this:

Number and Percent of Students Earning Each Integer Score Value

	1	**2**	**3**	**4**	**5**	**6**	
n =	1587.5	5937	33134	107821	39647	5210.5	193336
percent	0.8%	3.1%	17.1%	55.8%	20.5%	2.7%	100%

In this set, the 3 scores now include half of what would have been the 2.5's and half the 3.5's, and the 4 scores now include half of what would have been the 3.5's and half the 4.5's.

Using the Average of 3-and-above and 4-and-above

With only integer score values, the currently proposed procedure for estimating the percent scoring 3.5 or higher would be to average the percent scoring 3 or higher (96.1%) and the percent scoring 4 or higher (79%). For this set of data we would calculate [(96.1% + 79%) / 2] = 87.5%. This estimate is considerably higher than the actual value of 85% that we know existed when the full set of half-scores were available. The reason it is higher is that the new integer 3 score includes not only some of the previous 3.5's, but all of the old 3's and half the previous 2.5's. When this is averaged in with the 4 or higher percentage, the result tends to be slightly off the mark. For those interested in the exact arithmetic, it can be shown that the amount of distortion, in original score values, is equivalent to:

(.25*percent$_{2.5}$ + .5*percent$_{3}$ - .25*percent$_{3.5}$) which, for the current data is

(2.5 * 2.3% + .5 * 10% - .25 * 12.1%) = 2.5%

This is why the estimate based on a simple average of the 3-and-above and 4-and-above scores (87.5%) is 2.5% higher than it should be. This kind of formula would apply to all cases in which the half-value scores were represented only by integer-value scores. The unwanted additional value could result in a large positive distortion, a small amount of distortion, or even a negative distortion, depending on the original half-value score percentages.

A Better Formula for Estimating 3.5-and-above

The simple averaging approach for estimating the 3.5-and-above percent applies equal weight to the 3-and-above and the 4-and-above values found in the integer-only representation of the scores. As it turns out, equal weighting is not the best formula. There are proper differential weights for the 3-and-above and 4-and-above portion of the estimate that are calculable from the original data. The appropriate weight for the 3-and-above percentage is

[percent$_{3.5}$ / (percent$_{2.5}$ + 2 * percent$_{3}$ + percent $_{3.5}$)] which, for the current data is

= [12.1 / (2.3 + 2 * 10 + 12.1) = (12.1 / 34.4) = .35

and, the proper weight for the 4-and-above percentage is one minus that weight calculated above, that is 1 - .35 = .65. For the current example, we would estimate the 3.5-and-above percent as

$$(.35 * 96.1\%) + (.65 * 79\%) = 85\%$$

which exactly duplicates the original 3.5-and-above percentage.

Comparing the Formulas

If we apply the old equal-weighting strategy and the new differential-weighting strategy to each district's writing test results from 2009, we can get a good sense of the degree of improvement in estimating the true 3.5-and-above percentage by the new methodology.

Frequency of Estimate Errors

	4th Grade		8th Grade		10th Grade	
Estimate Error	Old Method	New Method	Old Method	New Method	Old Method	New Method
16					1	
15					0	
14					0	
13					0	
12					3	
11			1		1	
10	2		0		4	
9	0		0		3	
8	3		1		11	1
7	9		3		6	0
6	11		5	1	22	1
5	16	1	7	1	43	3
4	35	4	8	0	30	4
3	43	9	18	3	40	8
2	47	18	66	8	21	16
1	20	37	69	15	11	41
0	11	83	19	76	5	65
-1	3	37	3	73		45
-2	1	7	1	20		14
-3		4		3		2
-4		1		1		0
-5						1

In this table we have calculated estimates of the 3.5-and-above percentage from simulated integer-only data for each of the three writing prompts at each grade level for all districts in the state. We employed the same weights (.35 and .65) estimated from state-level data to all districts and grade levels.

It is easy to see that the simple average "old method" is biased toward higher percentages and imparts greater error than the new method based on differential weights. Even though the weights were generalized to each district from state-level data, the differential weighting method resulted in considerably more accurate estimates.

Discussion

The use of a single judge for each essay in the FCAT writing test in 2010 means that student scores can only take on integer values. It will no longer be possible to summarize a school's writing performance in terms of the percentage of students scoring 3.5 or higher. To maintain consistency with the traditional system of grading schools, some method of estimating 3.5-or-higher percentages from integer data must be employed.

There can be no debate - estimating the 3.5-or-higher percentage by simply averaging the 3-and-above percent with the 4-and-above percent will result in considerable bias. When this approach is applied to the smaller sample sizes of individual schools, the degree of potential bias would be even greater than that observed in this analysis of district summaries. It would undoubtedly result in assigning some schools with inappropriate school grades.

A more accurate estimate of the 3.5-or-higher percentage can be made with differential weights applied to the 3-and-above and 4-and-above percentages. Although only a single scorer will be used for each essay in the 2010 writing test, a 20 percent sample of essays will be scored by two judges for purposes of establishing reliability. From this two-judge sample, it will be possible to approximate the proper values for the weights to be applied in the 3.5-or-higher estimates. Even generalized weights from district-level data using the formulas presented in this paper will result in considerably more accurate estimates of school performance.

All reports distributed by Research Services can be accessed at http://drs.dadeschools.net.