Establishing a Value Table for Alaska

Richard Hill
May 9, 2005

Background

Alaska is considering using a measure of student growth from year to year as a substitute for the "safe harbor" provision of No Child Left Behind. Safe harbor measures the change in the percentage of proficient students in the same grade from year to year—that is, it tracks school progress across different cohorts of students. Alaska's proposed measure would parallel safe harbor, but measure the progress of students within a cohort from one year to the next. In addition, whereas safe harbor judges schools on a "proficient/not proficient" basis, the measure that Alaska proposes would give schools credit for students who are not yet proficient, but are on track to becoming proficient.

The Center for Assessment has proposed a system of "value tables" to measure the progress of students from one year to the next. In this system, a school is awarded a certain number of points depending on each student's performance level on any given year, when compared to that student's performance level the previous year. Alaska proposes to employ that system to measure student progress.

Alaska places students into four performance levels: Not Proficient, Below Proficient, Proficient and Advanced. For purposes of NCLB, Proficient and Advanced are combined into one category. To more precisely track non-proficient students' progress across years, the bottom two performance levels each will be divided in half, yielding the following performance levels:

- Not Proficient Minus
- Not Proficient Plus
- Below Proficient Minus
- Below Proficient Plus
- Proficient or Advanced

A key step in creating a value table is determining which changes in performance levels are more valued than others. Obviously, it is better for a student to remain at Below Proficient Minus from one year to the next than to decline to Not Proficient Plus; but is it better, for example, for a student to progress from Not Proficient Plus to Below Proficient Minus as it is from Below Proficient Minus to Below Proficient Plus? Each change involves the movement of a student up one level from the previous year; are both valued equally or is one to be valued more than the other?

Alaska's Assessment and Accountability Advisory Panel made an initial attempt at answering these questions during its May 4, 2005 meeting. After organizing into four small groups of 2-4 members each, the committee members ranked every possible pair of performance level combinations. This paper is a report of the process they followed and the results.

The Process

Staff from the Center for Assessment started by providing the committee with an overview of the process of using value tables to measure student growth from year to year. The panel learned of the

plans to propose using measurement of student growth as a substitute for NCLB Safe Harbor, to combine the upper two performance levels into one for purposes of NCLB accountability, to divide the two lower levels into halves in order to better measure student progress, and to award differential points to schools based on the changes in students' performance levels from one year to the next. The committee was in general agreement that this approach seemed plausible and more consistent with NCLB goals than the current safe harbor system. They understood that their assignment was to rank each of the cells in the following table from most desirable to least desirable:

Year 1 Performance Level	Year 2 Performance Level				
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced
Not Proficient Minus					
Not Proficient Plus					
Below Proficient Minus					
Below Proficient Plus					
Proficient or Advanced					

In order to facilitate the process of making the rankings, each small group of committee members received a set of cards. Each card had the text of one of the 25 cells written on it (e.g., one card read, "Year 1: Not Proficient Minus, Year 2: Not Proficient Minus"), and the 25 cards in the pack contained all possible combinations of results across years.

The committee noted that the value of each of the changes would be partially dependent on the grade the student was in when the change occurred. For example, it would be less valuable to see a student change from Not Proficient Minus to Not Proficient Plus if the student were in the tenth grade (when they must be proficient on the High School Graduation Qualifying Examination) than if the student were in the fourth grade (and therefore likely on track to becoming Proficient by the tenth grade). Rather than have the groups try to create the rankings for every possible grade, they were told to consider changes from grade 5 to grade 6.

The groups spent about half an hour making the initial sort. After that, the results of all groups were shown and thoroughly discussed. Several issues came to light, including the following:

1. One group placed additional value when students became Proficient or Advanced.
2. Another group placed additional value on students being at least Below Proficient Minus by the end of the sixth grade, feeling that students who were below that level at that point in their school careers likely would not be able to progress to Proficient by the tenth grade.
3. The groups spent most of their time sorting results along the diagonals. For example, there are four cells in the table that reflect students advancing one level from the previous year. All of the groups rated those four cells as more valuable than any of the five cells that reflect students remaining at the same level they were at the previous year. Choosing how to rank out those four cells, then, became the more difficult task for them to accomplish.

4. Within any diagonal, however, it was clear that scoring Not Proficient Minus needed to be awarded the fewest number of points, since the range for those scores could go all the way down to students turning in a blank test booklet. If a student was Not Proficient Minus one year and then Not Proficient Minus the next, it is possible that the student had received a zero on the test both years—a result that is hard to call "year-for-year" growth. Such a result could not be as valuable as scoring, for example, Below Proficient Plus two consecutive years.

After discussion, the groups reconsidered their rankings and made changes they deemed appropriate. To facilitate the reporting of results, we numbered each of the cells as follows:

Year 1 Performance Level	Year 2 Performance Level				
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced
Not Proficient Minus	1	2	3	4	5
Not Proficient Plus	6	7	8	9	10
Below Proficient Minus	11	12	13	14	15
Below Proficient Plus	16	17	18	19	20
Proficient or Advanced	21	22	23	24	25

Table 1 reports the rankings of the cells by each of the groups, from least desirable to most desirable. Both the initial and final rankings for each group are shown.

Table 1

Rankings of the Cells by Each of the Panel Groups

Ranking	Initial Round				Final Round			
	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
1	21	21	21	21	21	21	21	21
2	16	16	22	22	16	16	16	16
3	22	22	16	16	22	22	22	22
4	11	11	23	23	11	11	11	11
5	23	17	11	11	17	17	17	23
6	17	23	17	17	23	23	23	17
7	6	6	24	24	6	6	6	24
8	12	12	12	12	12	12	12	12
9	18	18	6	18	18	18	18	18
10	24	24	18	6	24	24	24	6
11	1	1	1	1	1	1	1	1

12	7	7	7	7	7	7	7	7
13	19	13	13	13	19	13	13	13
14	25	19	19	19	25	19	19	19
15	13	25	25	25	13	25	25	25
16	2	2	14	20	2	2	14	8
17	14	8	2	14	14	8	2	14
18	20	14	20	8	20	14	20	20
19	8	3	8	2	8	3	8	2
20	3	9	15	15	3	9	9	15
21	15	20	9	9	9	20	15	9
22	9	4	3	3	15	4	3	3
23	10	15	10	10	10	15	4	10
24	4	10	4	4	4	10	10	4
25	5	5	5	5	5	5	5	5

Converting the Results to a Value Table

Since this is the first time we have asked a committee to rank performance level changes this way, there is no prescribed way for converting their rankings to a value table. As an initial step, we computed the average final ranking for each cell in the table. Those results are provided in Table 2.

Table 2

Average Rank for Each Cell After Final Round of Panel Ratings

Year 1 Performance Level	Year 2 Performance Level				
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced
Not Proficient Minus	11	17	20.5	23.25	25
Not Proficient Plus	8	12	17	20.75	23.5
Below Proficient Minus	4	8	13.5	17.75	21.5
Below Proficient Plus	2	5.25	9	13.75	18.75
Proficient or Advanced	1	3	5.75	9	14.75

One initial thought was to multiple all the values in the table by 10 and round them off to the nearest 10. That certainly would be a value table that directly reflected the judgments of the panel. However, there were some disadvantages to doing that. First, Department staff had made a decision that scoring "Not Proficient Minus" in the second year should receive a value of 0, regardless of how the student had done the first year, since a student could turn in a blank paper and receive that performance level; i.e., it was a performance level that did not necessarily reflect any improvement from the previous year. Second, the average score across all the cells in the table would be, by

definition, 125; that is a score that does not have any inherently obvious interpretation (the way a score of, say, 100 would have). But most importantly, the rankings of the committee certainly were done thinking of students' true scores (e.g., a student being truly Proficient or Advanced one year and remaining truly Proficient or Advanced the second year), while the results of testing will reflect students' observed scores. The observed scores certainly are well correlated with the true scores, but there is little doubt that the observed scores for some students will not reflect their true change from one year to the next. It is certain the regression to the mean needs to be considered in the interpretation of the results. Students who score well below the state average one year will, on average, improve the next year, while students who score well above the state average one year will have, on average, a decline the next year. This regression effect needs to be taken into account in any evaluation of the proposed value table.

Since the statewide testing program does not yet test students at every grade, we needed to estimate what the results would be if students were tested in two consecutive years and each student got the same average effectiveness of instruction. When applied to a proposed value table, that would tell us what the distribution of scores would be under the condition of constant, "year for year" growth for each student. In order to make those estimations, we started with a set of assumptions:

1. Student scaled scores each year would be distributed normally.
2. The percentage of students performing at each performance level would be constant across the two years, and the percentage of students at each performance level would be equal to the grade 6 results for 2004.
3. The correlation between students' observed scores across years would be 0.80. This assumes a reliability of .92 for the tests within any given year, and a correlation of .96 for student true scores across years.

The percentages of students at each performance level in Grade 6 in 2004 are shown in Table 3.

Table 3

The Percentage of Students at Each Performance Level in Grade 6 for 2004

Content Area	Performance Level		
	Not Proficient	Below Proficient	Proficient or Advanced
Reading	10.3	19.5	70.2
Writing	3.0	20.8	76.2
Math	20.6	14.8	64.6

If student performance is normally distributed, then Table 4 provides the z-scores that would divide the performance levels.

Table 4

Z-scores at the Performance Level Cuts

Content Area	Performance Level Cut	
	Not Proficient to Below Proficient	Below Proficient to Proficient or Advanced
Reading	-1.27	-0.53
Writing	-1.88	-0.71
Math	-0.82	-0.38

Given these cuts, we next divided each of the lower two performance levels into halves (e.g., to create "Below Proficient Minus" and "Below Proficient Plus"). For Below Proficient, this was done by dividing the range for the level in half. For Not Proficient, this was done by computing the average width of the Below Proficient Minus category (it was 0.39), and subtracting that amount from the cut score between Not Proficient and Below Proficient. That led to the performance level cuts reported in Table 5.

Table 5

Z-scores for the Expanded Performance Level Cuts

Content Area	Performance Level Cut			
	Not Proficient Minus to Not Proficient Plus	Not Proficient Plus to Below Proficient Minus	Below Proficient Minus to Below Proficient Plus	Below Proficient Plus to Proficient or Advanced
Reading	-1.66	-1.27	-0.90	-0.53
Writing	-2.27	-1.88	-1.30	-0.71
Math	-1.21	-0.82	-0.60	-0.38

Using the z-scores in Table 5 as the cut points between the performance levels leads to the approximate percentages of students at each performance level shown in Table 6.

Table 6

Percentage of Students at Each Performance Level
When Using the Cut Scores Listed in Table 5

Content Area	Performance Level				
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced
Reading	5	6	8	11	70
Writing	1	2	7	14	76
Math	11	9	7	8	65

To create a simulated set of data, a computer program was written that generated a random scaled score for a student in Year 1 (a random pick from the normal distribution with a mean of zero and a standard deviation of 1), and then under the condition that scores in Year 2 needed to correlate 0.80 with scores in Year 1, generated a random score for the student in Year 2. These scaled scores were then translated into performance levels for each year, using the performance level cuts provided in Table 5. Each computer run generated 10,000 simulated students.

One such run generated the following statistics:

Year 1 Mean	-0.01
Year 2 Mean	-0.01
Year 1 St. Dev.	0.99
Year 2 St. Dev.	1.00
Correlation between Year 1 and Year 2	.80

This run produced the cross-tabulation of performance levels for reading shown in Table 7.

Table 7

Number of Students at Each Performance Level for
One Round of 10,000 Simulated Students in Reading

Year 1 Performance Level	Year 2 Performance Level					
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced	Total
Not Proficient Minus	234	106	80	19	22	461
Not Proficient Plus	117	129	134	94	75	549
Below Proficient Minus	76	120	196	194	247	833
Below Proficient Plus	36	104	205	267	562	1174
Proficient or Advanced	23	80	231	527	6122	6983
Total	486	539	846	1101	7028	10000

There are some interesting results in Table 7. First of all, the simulated data provided percentages of students at each performance level each year closely match the percentages provided in Table 6. Therefore, for purposes of this study, simulating data for 10,000 students is plenty of data. Second, it can be seen that a fair percentage of the students change performance level from year to year, even though the model provides for a correlation of .96 of true scores between years. Note that about 5 percent of the students who had an observed score in the lowest category the first year had an observed score in the highest category the second year. Again, this is not real change for these students, but mostly a function of the measurement error associated with the test, and the consequent difference between observed scores and true scores. The amount of change is closely related to the width of the performance level; the vast majority of students who had observed scores at the highest level the first year remained at that level the second year—that performance level is, by far, the one with the greatest width. Finally, note the effect of regression. More students at the performance levels below Proficient or Advanced move up a level or more than move down (students in those four levels all are scoring below the state average, and therefore have a tendency to increase their scores from one year to the next); they are replaced at the lower levels with students who scored at the Proficient or Advanced level the first year but then moved down the second year. It is this random mixing of students from one year to the next that needs to be taken into account when translating the input of the committee (which is thinking of terms of *true* scores) to a value table that will work with observed scores.

Suppose we simply translated the votes of the committee as suggested earlier in this paper; take the average rank for each cell, round to the nearest whole number, and then multiply by 10, but with the additional proviso that zeroes are used for the value of receiving a level of Not Proficient Minus in Year 2. That would yield the value table shown in Table 8.

Table 8

An Initial Proposal for a Value Table

Year 1 Performance Level	Year 2 Performance Level				
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced
Not Proficient Minus	0	170	200	230	250
Not Proficient Plus	0	120	170	210	240
Below Proficient Minus	0	80	140	180	220
Below Proficient Plus	0	50	90	140	190
Proficient or Advanced	0	30	60	90	150

To check on the reasonableness of this potential value table, we generated three sets of 10,000 students each (one for each content area) and computed what the student-level results would be if the value table were applied. Table 9 provides some summary information for the particular random sample generated (note that repeating this process would produce similar, but not identical, results, since the 10,000 students were randomly generated).

Table 9

Results after Applying the Value Table in Table 8 to Three Sets of Randomly Generated Student-Level Data

Year 1 Performance Level	Average Score for Students in Year 2		
	Reading	Writing	Math
Not Proficient Minus	102	134	86
Not Proficient Plus	138	150	133
Below Proficient Minus	146	156	142
Below Proficient Plus	143	149	132
Proficient or Advanced	141	143	137
Mean	139	144	131
Correlation between students' scaled scores and the number of points they earn on the value table	.18	.06	.27

Notice in particular the differences between writing and math. In writing, more students are placed at the higher performance levels than in math. Even though the same value table is applied to both content areas, and even though the data for both content areas assumes no growth from one year to the next, scores are considerably higher for writing than they are for math. This is true not only for the overall mean, but for the mean within Year 1 performance level. Also, the overall mean across the three content areas is around 140. Given that there is no improvement, this is a value that would created difficulty in interpretation (that is, there seems to be no inherent meaning to a score of 140).

Both test problems can be solved by subtracting 40 points from each of the cells in the original value table, with two exceptions: first, no value in the table can be less than 0, and second, since the scores for the Not Proficient Minus students are already too low, subtract only 30 points from those values. Applying that rule yields the value table in Table 10, which is the recommended value table.

Table 10

Recommended Value Table for Alaska

Year 1 Performance Level	Year 2 Performance Level				
	Not Proficient Minus	Not Proficient Plus	Below Proficient Minus	Below Proficient Plus	Proficient or Advanced
Not Proficient Minus	0	140	170	200	220
Not Proficient Plus	0	80	130	170	200
Below Proficient Minus	0	40	100	140	180
Below Proficient Plus	0	10	50	100	150
Proficient or Advanced	0	0	20	50	110

Table 11 shows the statistics produced by this value table when applied to three randomly generated sets of 10,000 students.

Table 11

Results after Applying the Value Table in Table 10 to Three Sets of Randomly Generated Student-Level Data

Year 1 Performance Level	Average Score for Students in Year 2		
	Reading	Writing	Math
Not Proficient Minus	84	97	69
Not Proficient Plus	110	118	111
Below Proficient Minus	116	120	109
Below Proficient Plus	100	112	97
Proficient or Advanced	101	103	98
Mean	102	106	96
Correlation between students' scaled scores and the number of points they earn on the value table	.10	.03	.18

The statistics associated with this value table are an improvement in every category over the originally proposed one. First, the average scores for the three content areas are much closer to each other, both the overall average and the average for each Year 1 performance level. Second, the correlations between starting scaled score and the number of points earned by students is considerably lower (although higher for math than the other two content areas). Finally, the average across the three content areas is slightly above 100, so interpretation of the results will be simpler: an average score of 100 means students are making year-for-year growth, while a score above 100 means they are making more than year-for-year growth.