SAS® EVAAS® Technical Report

Ohio Scale Stabilization Process

May 14, 2010

Note from ODE: The Ohio Department of Education has reviewed its value-added measures from the perspective of how they can most effectively be used to inform instruction. Even though standard psychometric methods are used to provide for equivalent scales within a grade and subject, it is recognized that unanticipated variability emerges across grades and among years within a grade in the OAA scaling. This can impact the interpretation of value-added analyses due to the multi-year nature of the technique. ODE requested SAS develop an additional procedure within its analytical process to assure that grade- and subject- level results can be appropriately analyzed and interpreted. The procedure outlined below will provide the necessary stability in the value-added computations without compromising the integrity of either the procedures used by SAS or the underlying assessments.

1 Indications of Scale Instability

Ohio Achievement Test (OAT) scale scores are scaled so that the performance level of "proficient" corresponds to a score of at least 400 (cut scores for other performance levels vary by subject and grade). Statewide mean scale scores for the Spring examinations in each year are shown in Table 1[A]. Standard deviations vary somewhat but tend to be around 30. On such a scale, in a population as large as that of Ohio (with around 130,000 to 140,000 students per grade), one would generally observe only a very small statewide change in average scale score. The observed variation shown in Table 1 includes shifts that are much larger. Note, in particular, 5th grade Reading in 2008 and 2009, as compared to 2007, and 8th grade Reading in 2009, as compared to 2008.

		Grade					
Subject	Year	3	4	5	6	7	8
Mathematics	2005	414.9				408.8	408.8
	2006	418.4	421.7	407.4	416.2	410.3	412.5
	2007	424.3	420.1	409.2	423.1	414.0	412.8
	2008	420.8	419.6	408.6	424.7	417.0	418.4
	2009	422.4	425.0	410.6	425.1	418.2	415.8
Reading	2004	415.7					
	2005	414.4	420.4	417.5			420.5
	2006	415.1	419.7	422.0	426.6	421.4	421.1
	2007	415.5	424.6	421.5	419.0	418.9	420.7
	2008	414.0	421.6	**413.8**	422.9	418.5	422.2
	2009	415.4	427.8	**415.7**	424.1	418.9	**415.6**

Table 1: OAT State Mean Scale Scores

Unexpected shifts of this magnitude can also be reflected in percentages of students categorized as proficient (or better). Table 2 shows the statewide percent proficient[A]. Again, notice the especially large drops in 5th grade Reading in 2008 and 2009, and 8th grade Reading in 2009.

[A] Source: `http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=1143&ContentID=9479&Content=77476`

		Grade					
Subject	Year	3	4	5	6	7	8
Mathematics	2005	70.2				58.1	59.7
	2006	74.8	76.9	62.4	68.3	63.0	68.4
	2007	84.3	75.7	60.8	73.7	70.9	71.2
	2008	79.0	74.3	61.2	76.4	68.3	72.6
	2009	81.7	78.8	62.5	75.5	74.5	70.8
Reading	2004	75.1					
	2005	74.9	76.6	76.8			78.8
	2006	71.2	76.7	75.2	83.7	78.9	77.1
	2007	76.1	79.9	80.0	77.6	77.4	80.0
	2008	74.0	81.0	**72.2**	79.5	76.9	79.2
	2009	75.4	82.5	**72.3**	81.9	76.8	**72.6**

Table 2: OAT Percent Proficient

In the case of percent proficient, besides the variation from year to year within grades, there is also notable variation between grades. For example, the percent proficient in Mathematics in recent years (2007-2009) is around 80% in 3rd grade but just above 60% in 5th grade. This is also reflected in the mean scale scores with a mean Mathematics score around 420 in 3rd grade and only around 410 in 5th grade. This is no doubt a consequence of the standard setting process and is not, in and of itself, a problem. However, for value-added analyses, grade-to-grade variations such as those seen in the mean scale scores would make for problems in the interpretation of value-added estimates expressed as scale score gains since the amount of gain that is to be expected in any subject and grade can vary widely. It is for this reason that SAS EVAAS value-added modeling is done with normal curve equivalents (NCEs) based on the statewide distribution of scale scores in a selected base year (for Ohio, the base year is currently 2007). The statewide means of the NCE scores are presented in Table 3. NCE scores are scaled to have a mean of 50 and a standard deviation of approximately 21. In Table 3, the means in the base year of 2007 are, of course, all equal to 50. On the NCE scale, a statewide shift of even 1 or 2 points could have a dramatic impact on value-added results.

Using NCEs removes the grade-to-grade variation, at least in the base year, but it does not solve the problem of unanticipated changes in scales from year to year within the same grade. Since value-added estimates are expressed as gains, one can get an idea of the consequences of scale instability for value-added by looking at mean NCEs along the diagonals in Table 3. For example, the mean NCE in 2009 4th grade Mathematics is 53.1 versus a mean of 47.4 for the same cohort in the previous year and grade (3rd grade 2008). As depicted by the scales, this cannot merely be attributed to "improvement" in Mathematics ability. Similarly, the mean NCE in 2009 8th grade Reading is 46.2 versus a mean of 49.6 in the previous year and grade. Such a statewide drop in Reading ability, as depicted by the scaling, is highly unlikely.

When these shifts are observed, although present in only a handful of subjects and grades, they can have have a substantial effect on the value-added results. Figures 1 and 2 show the consequences of

		Grade					
Subject	Year	3	4	5	6	7	8
Mathematics	2005	42.8				45.8	46.6
	2006	45.4	51.0	48.9	46.2	46.9	49.7
	2007	50.0	50.0	50.0	50.0	50.0	50.0
	2008	**47.4**	49.6	49.6	50.8	52.1	54.4
	2009	48.4	**53.1**	50.9	51.1	53.1	52.3
Reading	2004	50.0					
	2005	48.9	47.0	47.1			49.8
	2006	49.0	46.5	50.7	55.9	52.0	50.6
	2007	50.0	50.0	50.0	50.0	50.0	50.0
	2008	49.5	47.6	44.7	53.0	**49.6**	51.3
	2009	50.5	52.1	45.9	53.9	49.9	**46.2**

Table 3: OAT State Mean NCEs Basis 2007

scale instability for value-added estimation for specific subject and grade combinations. Figure 1 shows that in 2009 4th grade Mathematics, the vast majority of schools were "plus" (shown as green) while only about one-third of the schools had been "plus" in the previous year. Figure 2 shows the instability even more substantially for Reading, with the percent of "plus" schools varying from near 0% in 2008 5th grade to near 100% in 2009 6th grade. The purposes of a standards approach for value-added is to assure that shifts in the overall population are appropriately recorded. At the request of the Ohio Department of Education, a procedure was developed and presented herein to dampen the variability in the OAT scale to ensure that any changes in scale reflect reasonable changes in the state's population so as to enhance the appropriate interpretation of the value-added results.

2 Scale Stabilization Procedure

Scale stabilization is accomplished by modifying the mean NCEs shown in Table 3 to prevent the abrupt changes seen in that table. The overall procedure is as follows.

- Stabilize the table of mean NCEs
  - Convert student scale scores to NCEs using the 2007 state distributions of scale scores for each subject and grade.
  - Obtain the table of mean base-year NCEs as shown in Table 3.
  - Adjust the table of means to stabilize cohort fluctuation. See below for details.
  - Subtract 50 from each stabilized mean NCE to obtain a table of corrections.
- Stabilize the student NCE scores

- Convert all student scale scores to within-year NCEs using the state distributions of scale scores for each subject, grade and year.
- Adjust the within-year NCE for each student by applying the value from the previously calculated table of corrections.

The heart of the procedure is the replacement of each mean NCE in Table 3 by a weighted average of that targeted mean with "adjacent" means. This prevents the targeted mean from deviating too far from its adjacent means. The weights used in the weighted averages are shown in Table 4. Specifically, each targeted mean (red) is averaged with two "adjacent" means that are closer to the base year. One of the adjacent means is in the same cohort (green; on the diagonal from the targeted mean); the other is in the same grade (blue; in the same column of the table). This operationalizes the idea that statewide means within the same cohort should not vary drastically from year to year, nor should statewide means in the same grade (but different cohorts) vary drastically from year to year. For value-added analyses, the cohort means are more important since the along-cohort changes are a reflection of gains. Therefore, the adjacent mean in the same cohort receives more weight (50%) than the adjacent mean in the same grade (25%). The targeted mean itself receives a weight of 25%. For 2008 and 2009 3rd grade means, there is no adjacent mean in the same cohort (2nd grade). The only available adjacent mean is in the same grade; this adjacent mean receives the entire adjacent-means weight of 75%. Likewise, for 8th grade means in years prior to the base year, there are no available 9th grade means, so the adjacent 8th grade mean receives a weight of 75%.

	Grade					
Year	3	4	5	6	7	8
2005						
2006	0.25					
2007	0.25	0.50		0.50	0.25	
2008					0.25	
2009						

Table 4: Templates for calculating weighted averages of two targeted means (red cells), one after the base year (7th grade 2008) and one before the base year (3rd grade 2006).

This weighting process is applied sequentially. That is, it is first applied to the years adjacent to the base year (2008 and 2006). Then it is applied to years 2009 and 2005 using the stabilized 2006 and 2008 means as the "adjacent" means, and so on. In rare instances (when the targeted mean differs drastically from its adjacent mean), the resulting stabilized mean may still deviate substantially from it neighbors. In such cases, the along-cohort change in the mean is not allowed to move upward by more than 1.0 NCE point or downward by more than 0.5 NCE points. It would be highly unlikely to observe such large changes in an entire state's population. And, given the focus on enhancing educational opportunities for students, any potential decrease in a state's test scores would likely be smaller than possible increases – thus the more restrictive bound on decreases. The results of applying the above procedure to the Table 3 means is shown in Table 5.

Here are a few example calculations for one cohort:

2008 7th grade Reading: $(0.25 \times 49.6) + (0.50 \times 50.0) + (0.25 \times 50.0) = 49.9$.
2008 8th grade Reading: $(0.25 \times 51.3) + (0.50 \times 50.0) + (0.25 \times 50.0) = 50.3$.
2009 8th grade Reading: $(0.25 \times 46.2) + (0.50 \times 49.9) + (0.25 \times 50.3) = 49.1 \Rightarrow 49.4$.

Note that the calculation of the 2009 8th grade mean utilizes the stabilized means of 49.9 for the previous year/same cohort and 50.3 for the previous year/same grade. In addition, note that since the 2008 7th grade stabilized mean was calculated as 49.9 and the 2009 8th grade mean was calculated as 49.1 (a drop of 0.8 NCE points), the 2009 8th grade mean is revised to reflect a maximum drop of 0.5 NCEs, and a mean NCE of 49.4 is used in place of 49.1.

		Grade					
Subject	Year	3	4	5	6	7	8
Mathematics	2005	49.2				48.9	49.1
	2006	49.0	50.2	49.7	49.1	49.2	49.9
	2007	50.0	50.0	50.0	50.0	50.0	50.0
	2008	49.5	49.9	49.9	50.2	50.5	51.0
	2009	49.2	50.5	50.2	50.3	51.0	51.1
Reading	2004	49.4					
	2005	49.2	49.2	49.6			50.1
	2006	49.7	49.1	50.2	50.5	50.5	50.1
	2007	50.0	50.0	50.0	50.0	50.0	50.0
	2008	49.9	49.5	49.5	50.7	49.9	50.3
	2009	50.0	50.3	49.0	50.5	50.3	49.4

Table 5: Stabilized OAT Mean NCEs Basis 2007

As noted above, individual student NCE scores are obtained by converting scale scores to within-year NCEs and then adjusting them using corrections based on the stabilized mean NCEs of Table 5. These corrections are obtained by subtracting 50 from each stabilized mean NCE and are shown in Table 6.

Statewide mean within-year NCEs, by definition, have a value of 50 (therefore no table of these means is displayed). The within-year NCEs will also have a statewide standard deviation of approximately 21. Thus, the conversion of scale scores to within-year NCEs completely stabilizes not only the means but also the standard deviations of the scores. (The original scale scores vary from year to year and from grade to grade, not only in their mean scores as shown in Table 1, but also in their standard deviations.) In fact, within-year NCEs provide the maximum amount of scale stabilization, but they are less than ideal for value-added analyses because they provide no possibility for the state as a whole to show progress. When the corrections in Table 6 are applied to each student's within-year NCEs, the resulting "final" stabilized student NCE scores will have statewide means equal to those shown in Table 5. The statewide standard deviations will be approximately 21.

Figures 3 through 8 show results, like those in Figures 1 and 2, from school-level value-added analyses using three different scalings. The "BaseYr" scaling uses the 2007 base-year NCEs that

		Grade					
Subject	Year	3	4	5	6	7	8
Mathematics	2005	−0.8				−1.1	−0.9
	2006	−1.0	+0.2	−0.3	−0.9	−0.8	−0.1
	2007	0.0	0.0	0.0	0.0	0.0	0.0
	2008	−0.5	−0.1	−0.1	+0.2	+0.5	+1.0
	2009	−0.8	+0.5	+0.2	+0.3	+1.0	+1.1
Reading	2004	−0.6					
	2005	−0.8	−0.8	−0.4			+0.1
	2006	−0.3	−0.9	+0.2	+0.5	+0.5	+0.1
	2007	0.0	0.0	0.0	0.0	0.0	0.0
	2008	−0.1	−0.5	−0.5	+0.7	−0.1	+0.3
	2009	0.0	+0.3	−1.0	+0.5	+0.3	−0.6

Table 6: Corrections to within-year NCEs

have been used in Ohio value-added analyses up to the present. The statewide means of these NCEs were shown in Table 3. The "IntraYr" scaling uses the within-year NCEs without the corrections from Table 6. The statewide means of these NCEs are all equal to 50. This is equivalent to re-norming the data each year, thereby removing any information about statewide trends over time. The "StableVA" scaling uses the final, stabilized NCEs obtained by adding the corrections from Table 6 to the within-year NCEs. The statewide means of these NCEs were shown in Table 5.

As would be expected, the use of within-year ("IntraYr") NCEs results in roughly the same number of "plus" (green) schools as "minus" (red) schools. Using base-year NCEs, on the other hand, results in certain year/grade/subjects with many more "minus" schools than "plus" (see 2007 7th Reading) and some with many more "plus" schools than "minus" (see 2009 6th Reading). The approach outlined here (labeled "StableVA" on charts) maintains a reasonable distribution of value-added results at the subject/grade level, while simultaneously reflecting changes in the state's performance over time.

Note that this scale stabilization procedure has been described in the context of the value-added analyses currently provided to all schools in the state of Ohio. These analyses assumed a base year of 2007. It should be noted, however, that this approach could be used with any selected base year, with scales adjusted in a similar manner, starting with the base year and working outward.

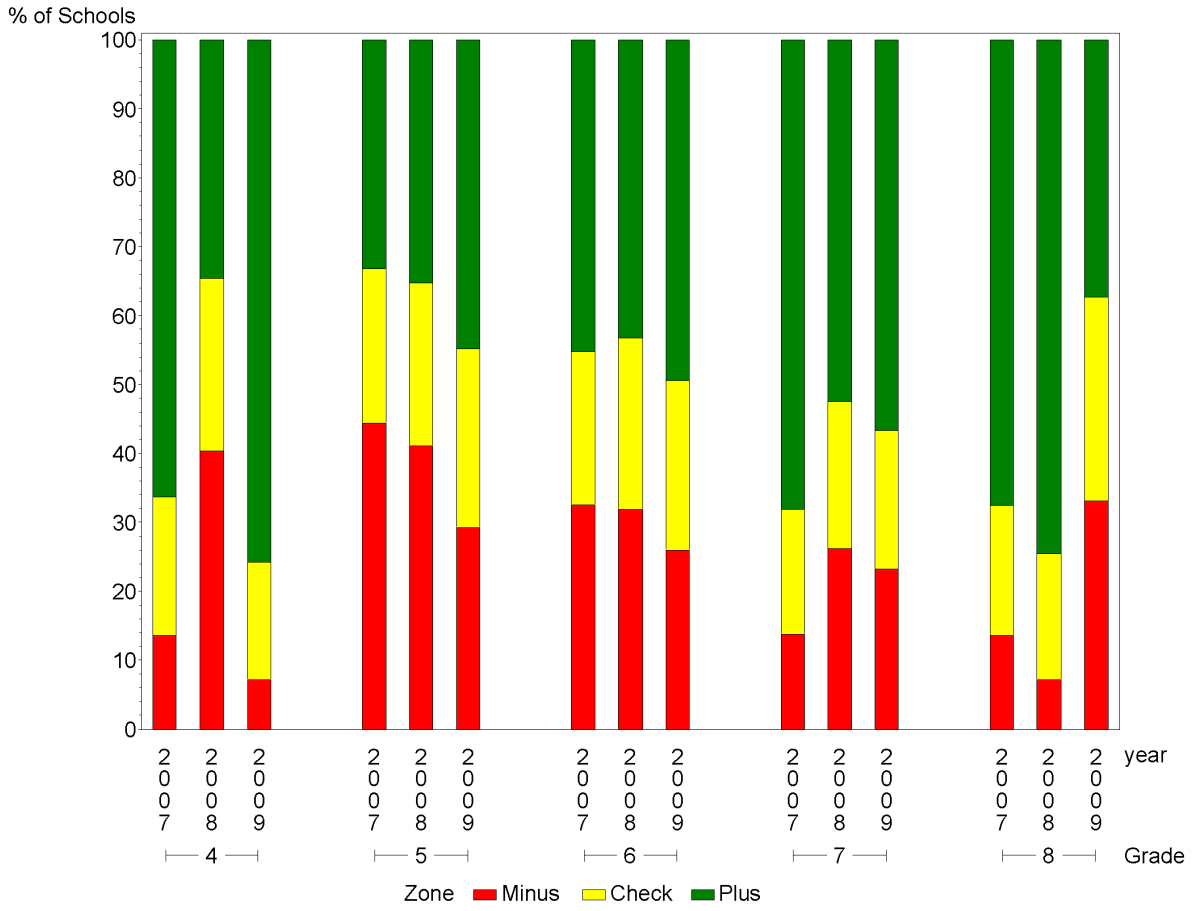

Figure 1: Percentage of schools in each value-added category for OAT Mathematics using base-year NCEs.

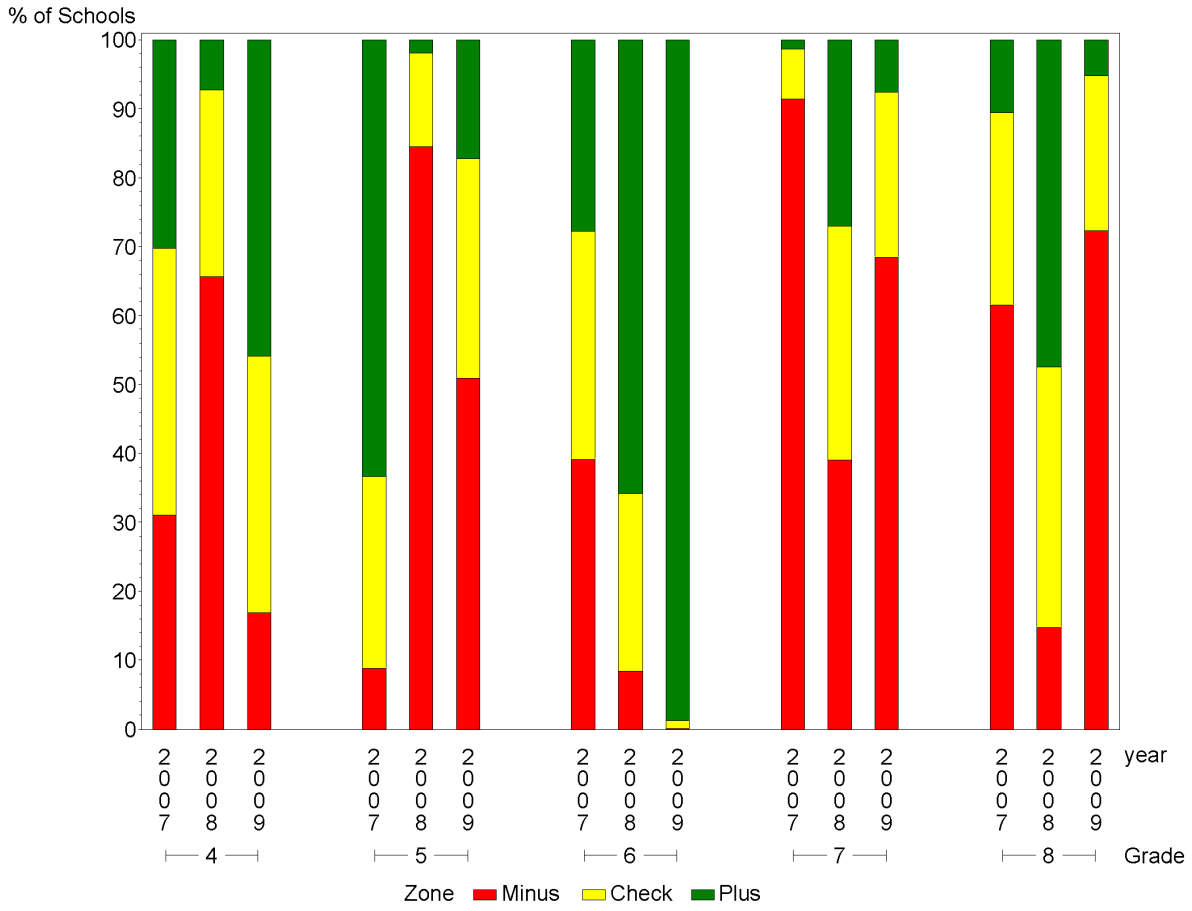

Figure 2: Percentage of schools in each value-added category for OAT Reading using base-year NCEs.

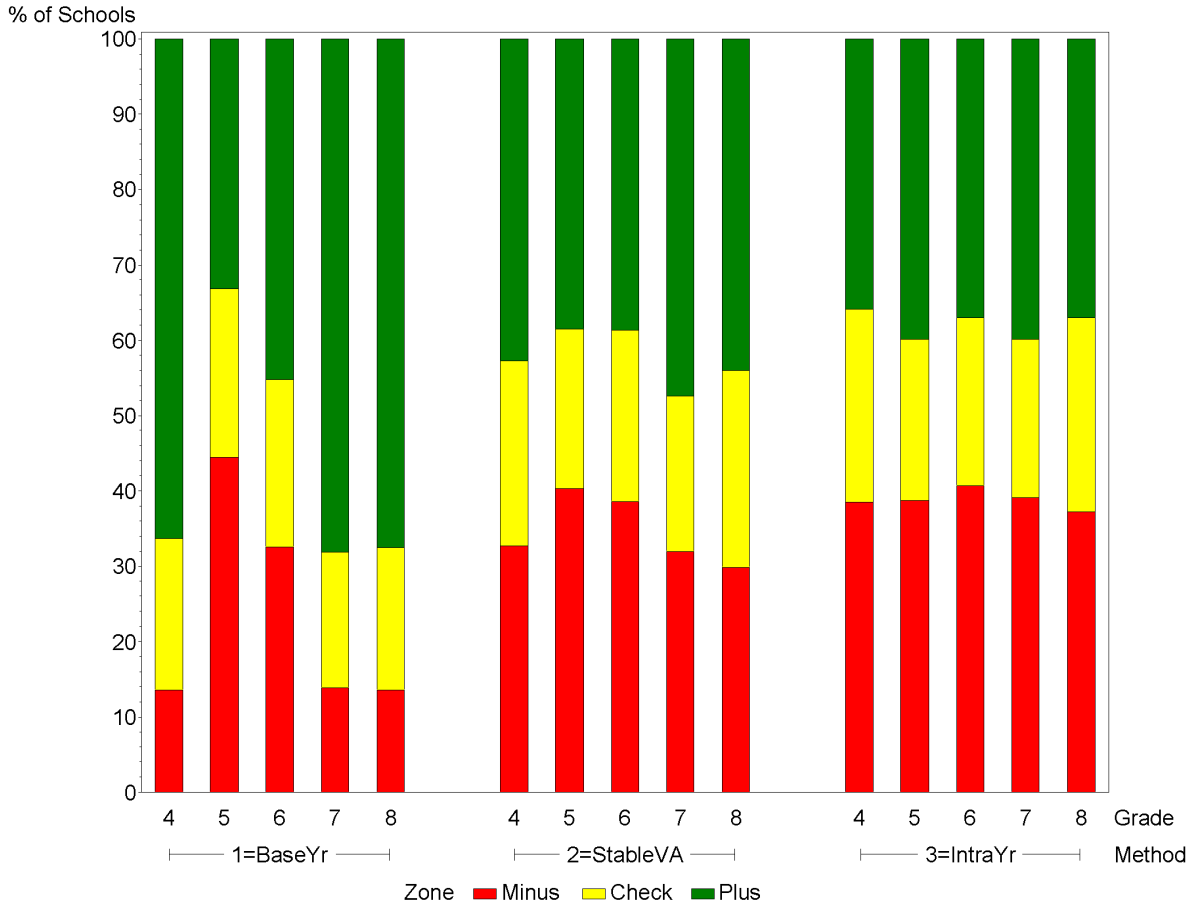

Figure 3: Percentage of schools in each value-added category for 2007 OAT Mathematics using three different scalings.

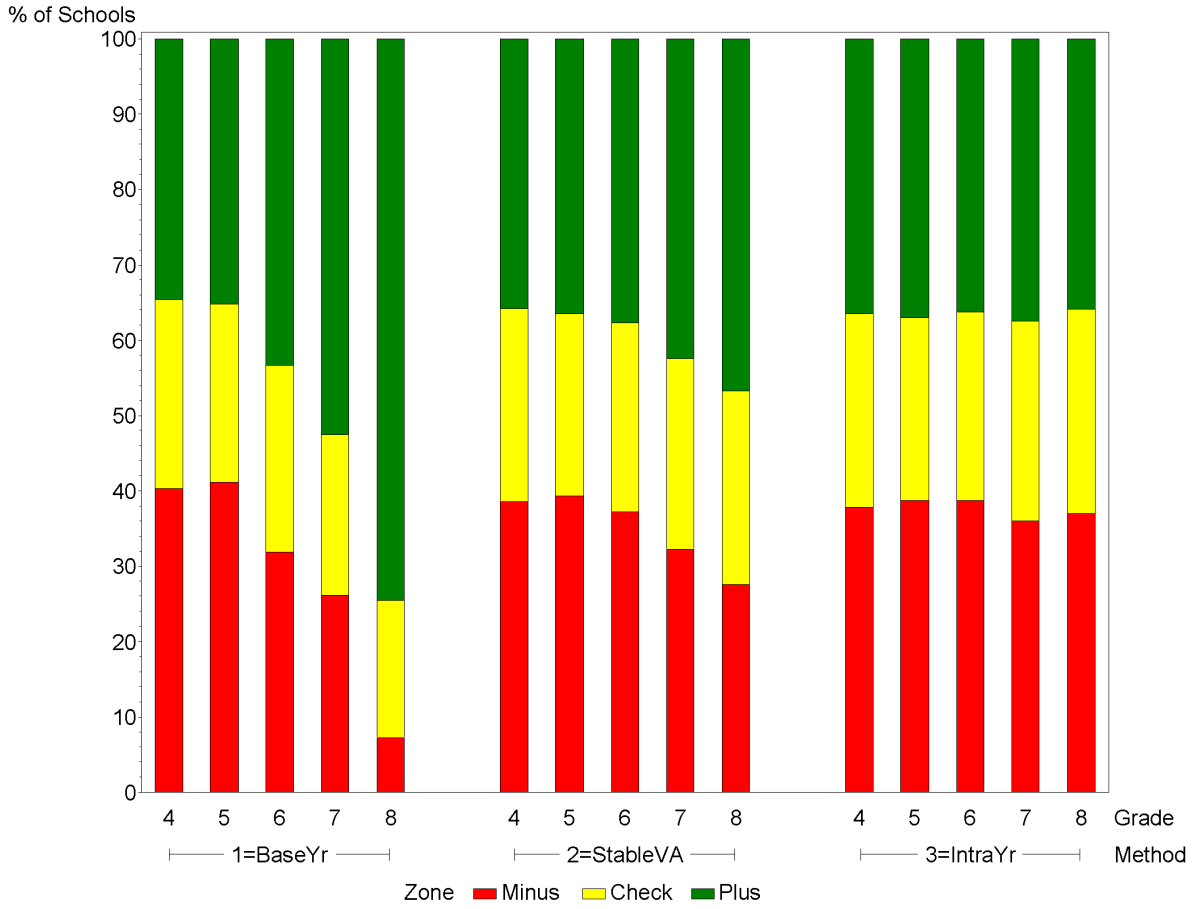

Figure 4: Percentage of schools in each value-added category for 2008 OAT Mathematics using three different scalings.

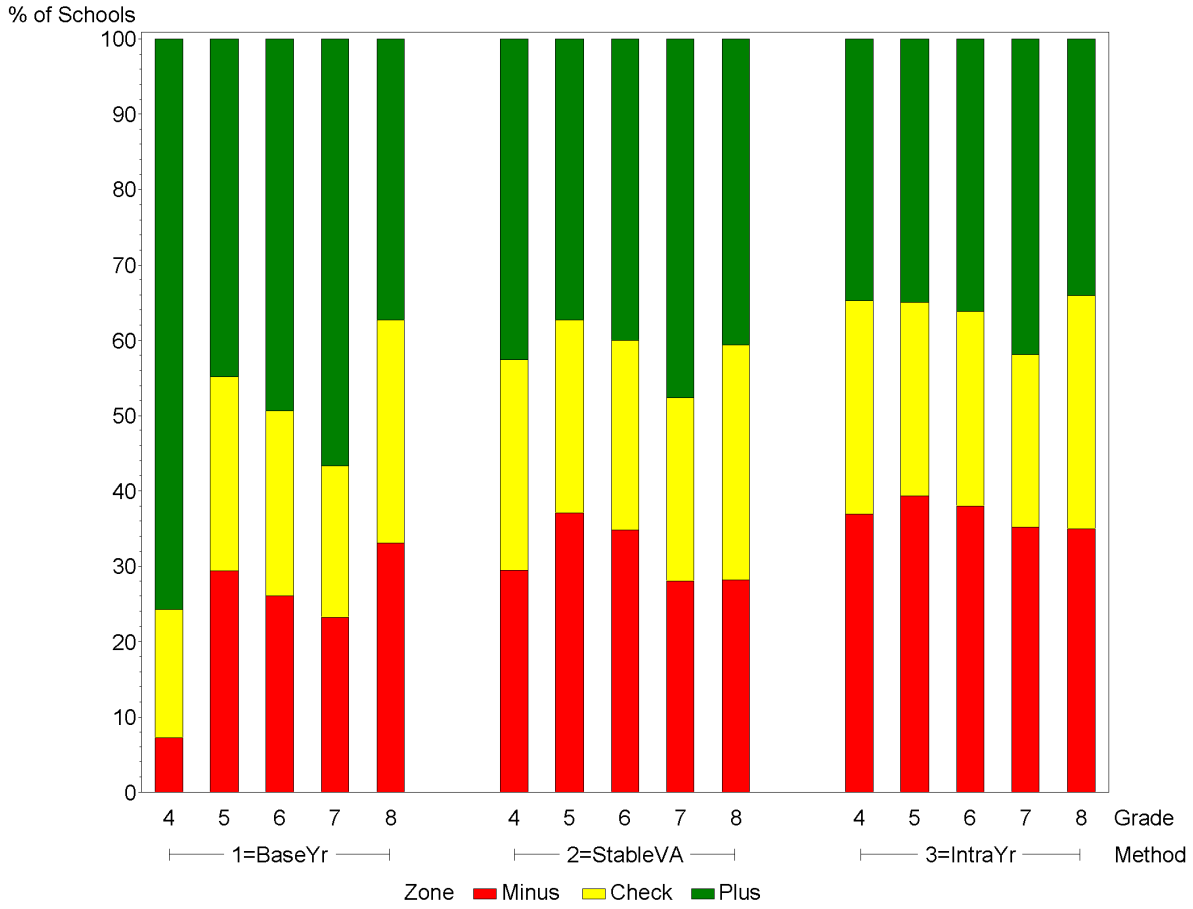

Figure 5: Percentage of schools in each value-added category for 2009 OAT Mathematics using three different scalings.

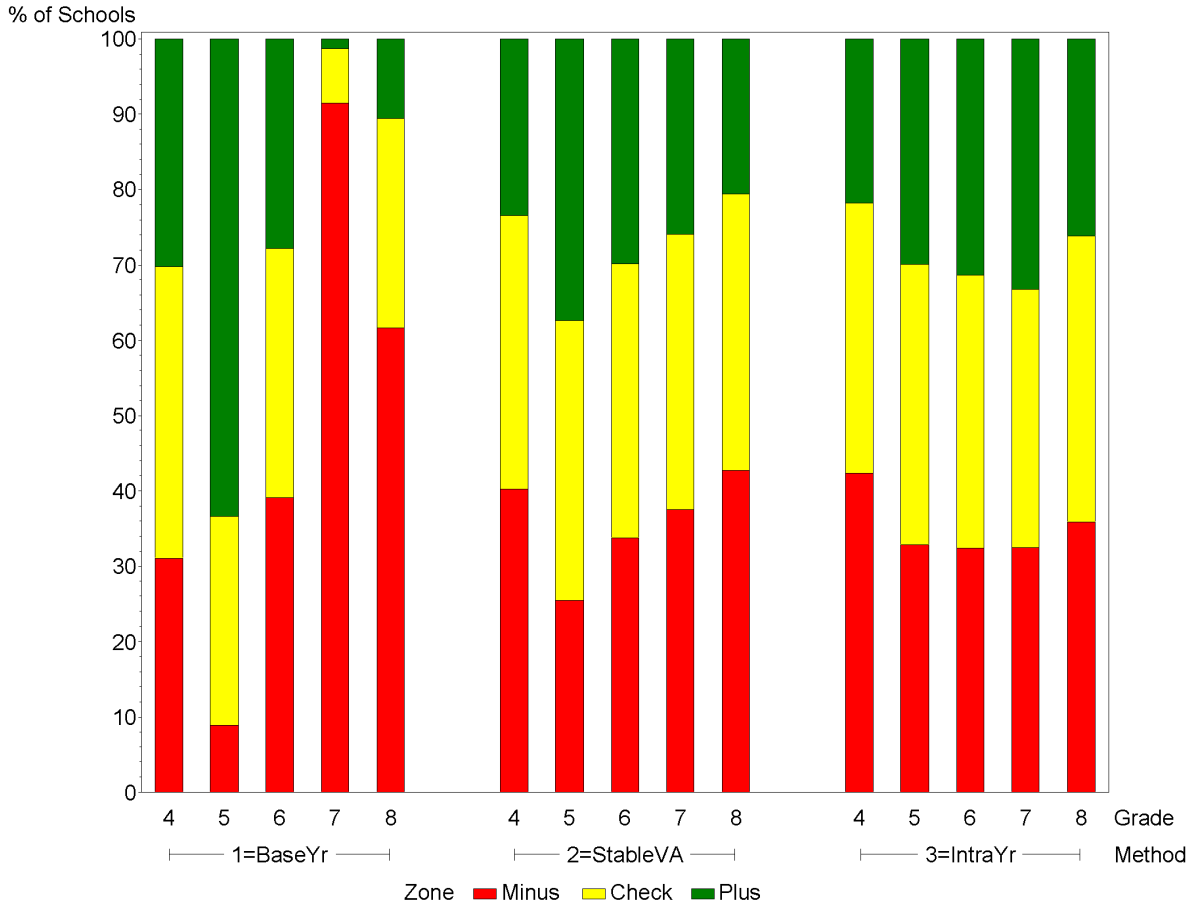

Figure 6: Percentage of schools in each value-added category for 2007 OAT Reading using three different scalings.

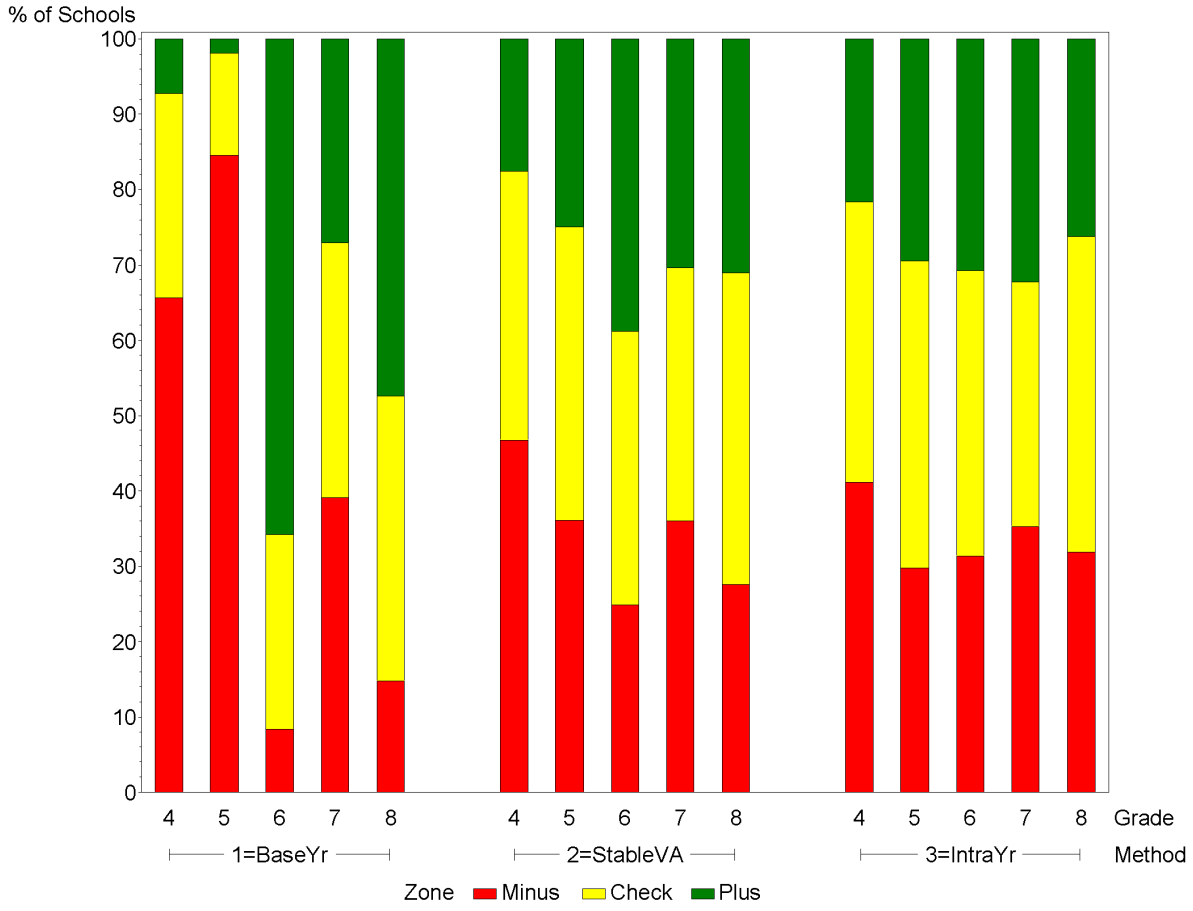

Figure 7: Percentage of schools in each value-added category for 2008 OAT Reading using three different scalings.

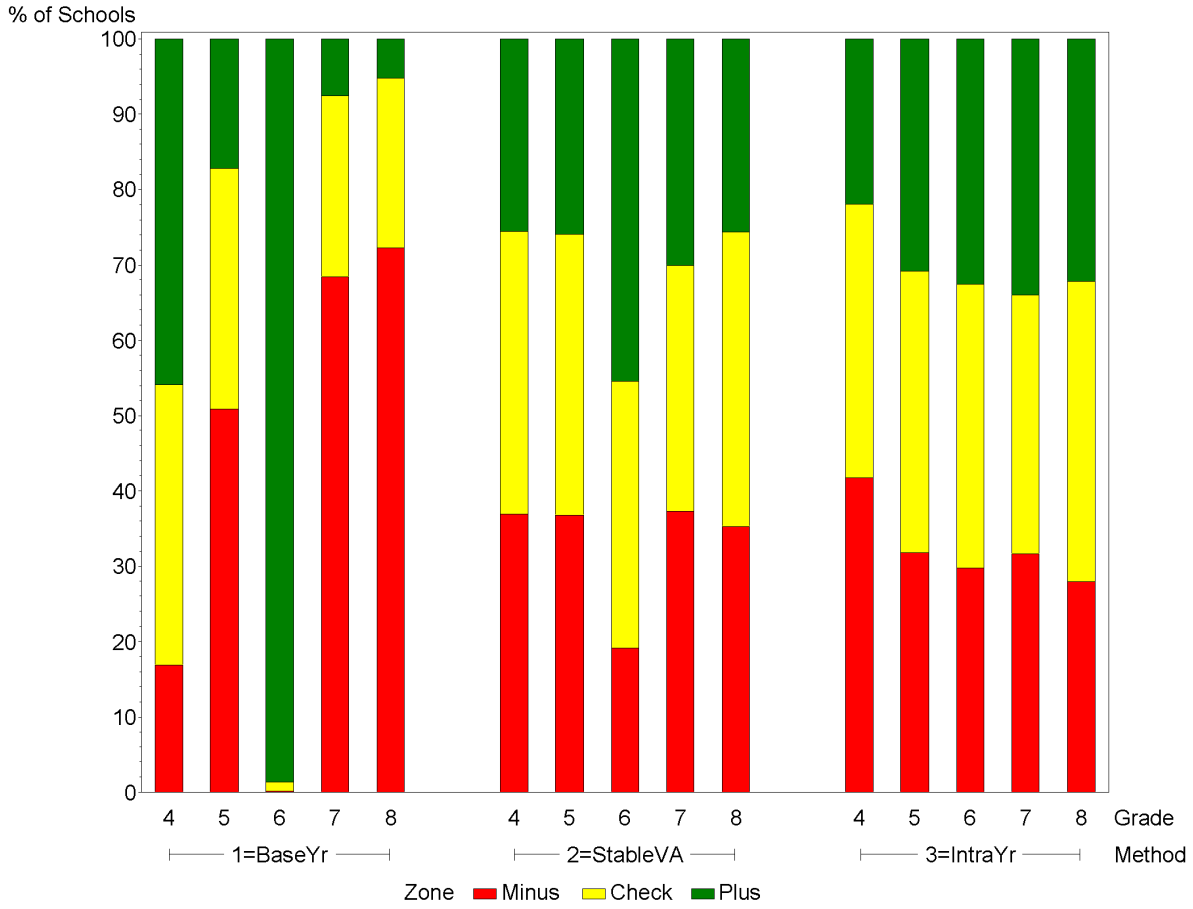

Figure 8: Percentage of schools in each value-added category for 2009 OAT Reading using three different scalings.