

COMPARING DIFFERENT ACCOUNTABILITY MEASURES: STATUS, IMPROVEMENT, INDEX, GROWTH—HOW ARE THEY ALIKE AND HOW DO THEY DIFFER?

Marianne Perie, Center for Assessment
John Weiss, Pennsylvania Department of Education

June 17, 2009

In November 2005, the U. S. Department of Education (USED) issued guidance for a pilot growth program and encouraged states to submit proposals for using growth models to demonstrate accountability under the federal *No Child Left Behind* (NCLB) Act. As part of the pilot, USED announced in 2005–2006 that it would approve no more than 10 high-quality growth models. More than 20 states submitted proposals and 9 states ultimately were approved for the pilot study. Then in December 2007, USED expanded the study to allow all states to submit a proposal for a growth model. Pennsylvania Department of Education (PDE) first submitted a growth model proposal in October 2006. After several revisions, they received conditional approval to use their growth model in January 2009.

Some issues that PDE faced during this process involved questions about differences between the growth model and their index. On June 10, 2008, Kerri Briggs (then-Assistant Secretary for Elementary and Secondary Education) sent a letter to Pennsylvania’s state superintendent denying approval of their request to incorporate a growth model in the state accountability system because of an “existing performance index.” In her letter Briggs asserts: “The heart of the peers concerns relates to the interaction of Pennsylvania’s existing performance index with the growth model. The Department has concerns about the appropriateness of allowing a state to include both a performance index and a growth model in its accountability system.”

Furthermore, guidelines released in December 2008 by USED through a letter from Briggs to chief state school officers regarding the use of indices in accountability systems included a requirement that an index may be used only in lieu of a percent proficient model, not in addition to it.

This paper addresses the differences in the goals and values behind each type of model used in Pennsylvania’s accountability system. Many of the presuppositions are not unique to Pennsylvania but generalize to these types of measures used in any system. Then, the paper will analyze the differences in schools identified as meeting the target under each measure.

Pennsylvania’s Current Accountability Model

Pennsylvania currently sets goals using a status model, but they also include a safe harbor (improvement) measure and an index in their accountability system.

The status measure includes both a confidence interval and a 2-year average, but since these are both statistical mechanisms used to minimize the risk of Type I errors (identifying a school as needing improvement when it actually met the goal), this paper will focus only on the measure of percent proficient, rolling those two adjustments into the overall discussion of status. Pennsylvania has four performance levels: Below Basic, Basic, Proficient and Advanced. Students must score at the Proficient level or above to receive credit for meeting the status goal.

Additionally, Pennsylvania adopted the traditional safe harbor mechanism of reducing the percentage of students below proficient by at least 10 percent each year. Again, a confidence interval is used as a step in this calculation but this paper will focus on the meaning behind the central improvement measure.

The third primary measure is the Pennsylvania Performance Index (PPI). The index is calculated for each school by multiplying the percentage of students scoring at each achievement level by a specified weight. For purposes of calculating an index sensitive to improvement below the Proficient category, PDE split the two categories below Proficient in half, creating a high and low category within the Basic and Below Basic levels. Table 1 shows the weights allocated for each performance level reached by each student within a school.

Table 1. Weights Used in Calculating the PPI

Advanced	1.0
Proficient	1.0
Hi Basic	0.8
Lo Basic	0.6
Hi Below	0.4
Lo Below	0.2
Non-tested	0.0

Then, the totals are summed to arrive at the index score. The goal is for all schools to have an index value of 1.0 by the end of school year 2013–2014. To calculate interim targets, the difference between the current year’s index and 100 is calculated. Then that difference is divided by the number of years between the current year and 2014. As an example, Table 2 shows a calculation for a fictitious school.

Table 2: Example of PPI Calculations for Random Elementary School

Performance Level	Number of Students	Percentage of Students	Weight	Multiple
Advanced	50	14.3%	1.0	14.3
Proficient	100	28.6%	1.0	28.6
Hi Basic	75	21.4%	0.8	16.8
Low Basic	50	14.3%	0.6	8.6
Hi Below Basic	50	14.3%	0.4	5.7
Low Below Basic	25	7.1%	0.2	1.4
Sum	350			75.4

If this school has an index value of 75.4 in 2008 and needs a value of 100 in 2014, then they need to improve their score by 24.6 over six years, or by 4.1 each year. Thus their interim targets would be 79.5 for 2009, 83.6 in 2010, 87.7 in 2011, 91.8 in 2012, 95.9 in 2013, and 100 in 2014. Note that this measure allows for the inclusion of all students in the calculation regardless of whether they took a general or alternate assessment.

Internally, PDE uses the Pennsylvania Value-Added Assessment System (PVAAS) as an additional source of information for their districts regarding student growth. Because value-added systems do not meet federal requirements for growth under NCLB, PDE used those data to create a projection to proficiency model for their federal growth model. The projection provides an alternate definition of proficiency from proficient based on status measures to proficiency based on projections of performance on a future PSSA test. These projections will be held to the same standards of percent proficient as the status measures (i.e., same AMO targets) to be consistent with the overall accountability system. For most grades, the projection to proficiency is within two years, except from grade 8 to 11 as there are no interim tests to use to measure growth.

Table 3: Present and Projected Grades

Present Grade	4	5	6	7	8
Projected to Proficiency in Grade	6	7	8	8	11

These four primary measures are summarized in Table 4. Note that while Pennsylvania does use confidence intervals and a 2-year average and therefore has more than four ways for each school to make its AMO, these other alternatives are simply statistical adjustments; there are four basic metrics, following four core philosophies used to judge the adequacy of each school.

Table 4. Different Components of Pennsylvania’s Accountability System

	Definition	Goals	Values	Theory of action	Question answered
Status	The percentage of students scoring at or above the Proficient cut score.	100% proficiency (or higher) by 2014	It is important to hold all students to the same criterion and measure the percentage reaching that criterion each year.	By setting one criterion for all students to meet schools will improve the achievement of all students and lower the performance gap among subgroups of students.	How good is the achievement of students this school?
Improvement (Safe Harbor)	The percentage of tested students in a school building, district, or subgroup below the proficient achievement minus the percentage from the preceding year.	Reduce the percentage of students below proficient by at least 10 percent each year.	It is important to recognize progress in bringing more students to the Proficient criterion even when the absolute goal of numbers of students at Proficient is not met.	By setting an intermediate bar, this system encourages schools who are well below the status criterion to continue to work to improve student achievement and bring more students to the Proficient bar.	Is the achievement level of this school improving?

Metric	Definition	Goals	Values	Theory of action	Question answered
PPI	The percentage of students scoring at each achievement level multiplied by the weight; sum the totals to get the index score.	PPI of 100 by 2014	It is important that progress be measured in a way that is sensitive to academic growth all along the achievement scale	Using an index will encourage schools to focus on raising the achievement of all students below Proficient, and not just focus on those who are right below the Proficient cut score.	Is this school raising the achievement of students below the Proficient cut score?
PVAAS – growth toward proficient	Using a longitudinal analysis of individual student achievement over at least 2 years, predicts the percentage of students on track to score at or above the Proficient cut score within two years	100% of students proficient (or advanced) or on track to be proficient (or advanced) by 2014.	It is important to recognize schools in which students have not yet achieved proficiency but have demonstrated significant growth towards proficiency. (Also identifies schools where students are proficient but declining.)	The use of this measure will encourage school districts to place at-risk students on an accelerated path to proficiency in both reading and math through the targeted use of resources, interventions, professional development and high standards.	Are the students in this school increasing their achievement at such a pace that they are on track to reach Proficient within two years?

As seen in Table 4, each component of their accountability system targets a different definition of a successful school. The status metric examines a school’s overall level of performance. Interim targets are set with the goal of reaching 100% proficient by 2014. The improvement metric compares the current cohort of students to the previous cohort to see if more students are becoming proficient over time. Then, the PPI still holds schools to reaching 100% proficiency by 2014 but also values improvement along the continuum below and to the proficient cut score. Again, the PPI analyzes cohorts of students and does not track individual students—it is still a status measure that raises the bar every year to force continuous improvement. The growth toward proficient model also values growth toward proficient, but does so by tracking the progress of individual students. It then calculates whether or not each individual student is on track to reach Proficient within the next two years. So, each model, with the exception of safe harbor (improvement) has a goal of 100% proficient and bases the interim targets on that goal. But, each model values a different type of improvement of a school or an individual student. PDE has determined that any school not making any target under any metric is truly in need of improvement. Using all four metrics allows PDE to target its resources across 3,105 schools.

When we map the Pennsylvania system to the criteria for defining a quality school shown in Table 5, we see that two of their metrics map to Model 1 (percent proficient and PPI), one maps to Model 2 (safe harbor), and one maps to Model 3 (growth toward proficient).

Table 5. Criteria for Defining a Quality School in an Accountability Model

Criteria of “Good”		
	Status	Change
Achievement <i>(in relation to standards)</i>	Model 1: How high do students in the school score on state assessments? What percentage of students meets the state standards?	Model 2: Is the school improving, or increasing, the performance of classes of students over time? Is the percentage of students meeting the state standards increasing from one year to the next?
Effectiveness <i>(in relation to past performance of students)</i>	Model 3: Are students learning as they progress through the grades? Are individual students making expected progress from grade to grade?	Model 4: Is the school becoming <i>more</i> effective—is it helping students (individuals, subgroups, or all) reach higher levels of achievement or learn relatively more over the years than was achieved or expected in the past?

Adapted from Carlson, D. (2002). Focusing State Educational Accountability Systems: Four Methods of Judging School Quality and Progress.

Although the U.S. Department of Education has questioned the use of a growth model in a system that already uses a performance index, it is important to recognize that an index is another method of status, not growth. The targets set for the index resemble the increasing targets set for percent proficient, with both measures resulting in a goal of 100% proficient by 2014.

School Identification Rates

In calculating AYP for each school and district, PDE applies the four models in a particular order. Currently, that order is status, improvement, PPI, and growth. Once a school or district has made AYP on one of the metrics, it is removed from the pool and not measured against the other accountability models. That is, if a school reaches its goal for percent proficient, PDE does not report whether it also made its improvement and PPI goals. However, the data are there to run these calculations.

Table 6 shows how the 3,105 schools in the state made AYP (or failed to make it) in 2008.

Table 6: AYP Results from 2008 Collapsing Confidence Intervals and Averages into Four Main Accountability Metrics

School That Made Performance Targets	Number of Schools	Percent of Schools
Made Targets Using Status Only (Without ANY safe harbor or PPI calculation in any subgroup)	1,407	45.31%
Made Targets Using Safe Harbor in At Least One Subgroup (But without any PPI calculation in any subgroup)	678	21.84%
Made Targets Using PPI in At Least One Subgroup	130	4.19%
Would Have Made Target via Growth (if it had been allowed in 2008)	232	7.47%
Total making target under these four methods	2,447	78.81%

However, if we reconsider the accountability model to look at each metric individually, the results differ. For instance, what if PDE were to ask the question: If I could only use one metric to determine whether or not each school was making adequate yearly progress, which metric would identify the most (or least) number of schools? Table 7 shows the results of calculating each target separately for each school, but requiring the school to make the target via that one metric for every subgroup in that school.

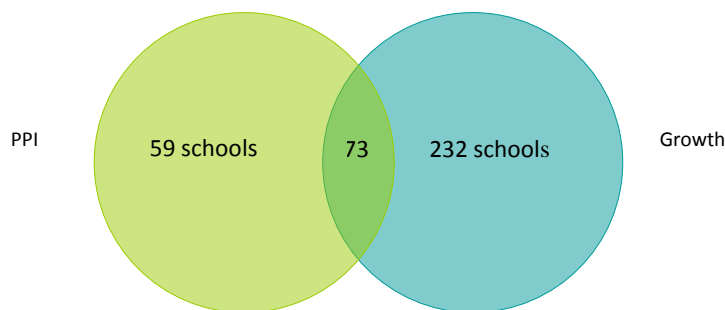
Table 7: Projected Results from 2008 Data on Using Each Metric Separately

School That Made Performance Targets	Number of Schools	Percent of Schools	Total Number of Schools in Calculation*
Using only percent proficient	1,407	45.48%	3105
Using only Safe Harbor (improvement)	361	12.12%	2978
Using only PPI	918	30.83%	2978
Using only Growth	2,145	72.03%	2978

*Schools must have scores for two consecutive years to calculate safe harbor, PPI or growth

Because the peer review comments seemed to imply that PDE would have to choose between the PPI and their proposed growth model, additional analyses were run comparing school identification under each metric. In 2008, of the 1,020 schools that did not make AYP using status or safe harbor, 132 would have made the PPI target and 305 would have made the growth target. However, these two numbers are not mutually exclusive. As shown in Figure 1, 73 schools made both the PPI and the growth targets, meaning 59 schools would make the PPI targets but not the growth targets, and 232 schools would make the growth targets but not the PPI targets. Clearly the different models are identifying different schools as being effective. It is important to understand the differences in these metrics.

Figure 1. Diagram showing the number of schools making PPI, growth, or both goals



Index and Improvement

Interestingly, there are more schools identified as making their target under PPI than under Safe Harbor improvement category. Perhaps this result can be explained by the fact that the PPI honors improvement across all levels of performance, not just at the Proficient cut score. For instance, consider the example school shown in Table 2 expanded here in Table 8.

Table 8: Random Elementary School Calculations for Two Years

Performance Level	Number of Students in Year 1	Percentage of Students in Year 1	Number of Students in Year 2	Proportion of Students in Year 2
Advanced	50	14.3	50	14.3
Proficient	100	28.6	105	30.0
Hi Basic	75	21.4	85	24.3
Low Basic	50	14.3	65	18.6
Hi Below Basic	50	14.3	40	11.4
Low Below Basic	25	7.1	5	1.4
Sum	350		350	
Percent Proficient		42.9%		44.3%
PPI	.754		.797	

From Table 8, we can see how a school that is below the target for percent proficient might meet their PPI target, but fail to meet the safe harbor goal. From the previous discussion, we know that to reach the PPI target of 1.0 in 2014, this school had to make an interim target of .795 in the subsequent year, which they did. At the same time, if they wanted to meet their safe harbor target—reducing the percentage of students not reaching Proficient by 10%-- they would have needed to reduce the percentage of students not reaching Proficient from 57.1% to 51.4%. However, the calculations in Table 7 show that the percentage of students reaching Proficient or above was 44.3%, meaning the percentage of students not reach Proficient was only reduced to 55.7%. Therefore, they did not meet their safe harbor target. An examination of the numbers show that most of the growth occurred from students moving from Below Basic to Basic, but very few students moved from Basic to Proficient. It is important to note that these “students” are not the exact same students from Year 1 to Year 2. Rather it includes the cohort of students who were in the school in Year 1 (e.g., all students in grades 3–5) compared to the cohort of

students in the school in Year 2, with no attempt made to match up the students themselves.

Growth to Proficiency Compared to Other Metrics

Because the growth model is a projection towards proficiency, it counts all schools that are currently meeting the percent proficient target as meeting their AMO. The only exception would be if students within a school are on a downward trajectory such that even though they are scoring at the Proficient level now, many of them are projected to be below Proficient within two years. Therefore most of the 1,412 schools included in the status results (such as those shown in Table 7) are included in the growth results as being on track to be proficient and meet the status goal. One of the newer requirements of USED is that an index should be used in lieu of percent proficient, not in addition to it.¹ Because this type of growth model includes all schools who are already proficient in its calculations, the only schools who would be lost by removing the percent proficient metric and replacing it with this growth model along with the PPI are those whose students are high achieving but losing ground.

The growth model identifies more schools as meeting the targets than the PPI, in part because it focuses on the growth of individual students and not the overall performance of two cohorts of students. Most likely, these schools include students that have moved from a lower “high basic” to a higher “high basic” moving them towards Proficient, but not in a way captured by the PPI. Conversely, the schools meeting their targets under PPI but not growth are most likely showing improvement from below Basic to Basic, but not at a rate that will allow enough students to reach Proficient within two years.

As a real example, consider Table 9, which shows an actual school from Philadelphia that made its AYP targets through the use of PPI and safe harbor. This is an inner-city school that is comprised primarily of African American students, a large proportion of whom are eligible for free or reduced-price lunch. Although not shown in this table, we know that in 2008 overall, 26% percent of students scored below basic, 27% scored at the basic level, 29% were proficient and 18% were advanced in math. In reading, 26% fell below basic, 25% at basic, 37% proficient, 14% were advanced. Comparing these numbers to those in 2007 showed that this school moved large numbers of students from below basic to basic between those two years. This type of growth is admirable and needs to be recognized, but requires the PPI to be identified. Because the improvement was at the lower end of the scale, this school would not have met the growth targets.

¹ As stated in the letter from Kerri Briggs to Chief State School Officers on January 2, 2009.

Table 9. Example of an Inner City Elementary School that Relied on PPI to Make AYP

	Academic Performance						Test Participation			
	Reading			Mathematics			Reading		Mathematics	
	% At/Above Proficient	Increase/Decrease from Last Year	Result	% At/Above Proficient	Increase/Decrease from Last Year	Result	% Tested	Result	% Tested	Result
2008 Targets:	63%			56%			95%		95%	
Students Overall	49.7	2.2	<u>PPI</u>	47.1	1.6	<u>PPI</u>	99.0	✓	99.8	✓
White non-Hispanic	-	-	-	-	-	-	-	-	-	-
Black/African American non-Hispanic	49.4	1.8	<u>PPI</u>	46.6	1.7	<u>PPI</u>	99.0	✓	99.8	✓
IEP-Special Education	45.9	5.9	<u>SHC</u>	44.3	11.6	<u>SH</u>	100.0	✓	100.0	✓
English Language Learners	-	-	-	-	-	-	-	-	-	-
Economically Disadvantaged	49.6	2.0	<u>PPI</u>	47.0	1.6	<u>PPI</u>	99.0	✓	99.8	✓

Table downloaded from PA Academic Achievement Report website: <http://paayp.emetric.net> on June 5, 2009.

Conclusion

In summary, it is important to articulate and recognize the different definitions for a successful school. Matching the definitions to the model and holding true to the goal of 100% proficiency by 2014 allows for the flexibility of using multiple metrics to calculate the effectiveness of a school. Ideally, the results of all metrics will be communicated to school leaders and analyzed at the state and district level to better understand where schools are doing well, where they are struggling, and any potential concerns for the future that may be identified early.

All of the metrics used in Pennsylvania are applied equally to all subgroups. The schools that do not meet any of the targets for each and every subgroup are considered the ones most in need of additional resources and attention in order to improve student outcomes.