4-7-2004

Education Policy Analysis Archives 12/14

Arizona State University

University of South Florida

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Volume 12 Number 14 **April 7, 2004** **ISSN 1068-2341**

How Feasible is Adequate Yearly Progress (AYP)? Simulations of School AYP "Uniform Averaging" and "Safe Harbor" under the No Child Left Behind Act

Jaekyung Lee
SUNY at Buffalo

Citation: Lee, J., (2004, April 7). How Feasible is Adequate Yearly Progress (AYP)? Simulations of School AYP "Uniform Averaging" and "Safe Harbor" under the No Child Left Behind Act. *Education Policy Analysis Archives, 12*(14). Retrieved [Date] from http://epaa.asu.edu/epaa/v12n14/.

Abstract

The No Child Left Behind Act of 2001 (NCLB) requires that schools make "adequate yearly progress" (AYP) towards the goal of having 100 percent of their students become proficient by year 2013-14. Through simulation analyses of Maine and Kentucky school performance data collected during the 1990s, this study investigates how feasible schools would have met the AYP targets if the mandate had been applied in the past with "uniform averaging (rolling averages)" and "safe harbor" options that have potential to help reduce the number of schools needing improvement or corrective action. Contrary to some expectations, the applications of both options would do little to reduce the risk of massive school failure due to unreasonably high AYP targets for all student groups. Implications of the results for the NCLB school accountability system and possible ways to make the

current AYP more feasible and fair are discussed.

The reauthorized Elementary and Secondary School Act (ESEA), No Child Left Behind Act of 2001 (NCLB), requires standards-based accountability for schools receiving Title I funds. One major component of this accountability policy is to report whether the schools are making "adequate yearly progress" (AYP) based on performance targets set by their state (i.e., 100% of students become proficient within 12 years from the baseline year). Since the passage of the NCLB, much concern has been raised about the AYP mandates and their possible consequences for schools that repeatedly fail to meet their AYP target (Linn, 2003).

Previous studies pointed out that some critical problems with AYP-based school accountability policies foreshadow technical challenges that lie ahead (Hill, 1997; Kane & Staiger, 2002; Kim & Sunderman, 2004; La Marca, 2003; Lee, 2003; Lee & Coladarci, 2002; Linn & Haug, 2002; Thum, 2002). While the studies raised technical issues such as reliability and validity with regard to AYP measures or pointed out policy implementation problems such as the lack of capacity and resources, the options available for schools to take advantage of under the NCLB have not been studied and discussed systematically. Specifically, there are two options available under the current NCLB legislation, that is, (1) uniform averaging (NCLB, 2001, Section 1111(b)(2)(J)) and (2) safe harbor (NCLB, 2001, Section 1111(b)(2)(I)), that might not only help improve the reliability or fairness of the AYP measure but also help save schools from failing to meet the AYP target. It remains to be examined whether and how those options might affect the feasibility of AYP that should be the most pressing issue for schools.

First, the uniform averaging procedure is designed to address a reliability issue: Does AYP measure schools' academic progress with sufficient consistency and stability? The typical school AYP measures tend to be highly vulnerable to fluctuation as they rely on comparison of successive cohort groups (as opposed to tracking the same cohort of students); it is particularly problematic in small schools which might have very few students for certain demographic category. In light of this difficulty, the NCLB permits aggregating data from multiple years to increase sample size for more reliable estimation of the target group's performance. While the term "uniform averaging" has not been clearly defined in either statistical or policy terms, it was interpreted as allowing for multiple approaches to aggregating multiple years' data and being able to use the techniques for either or both, status or/and improvement evaluations (Marion et al., 2002). For example, schools can average test scores from the current school year with test scores from the preceding two years, and this rolling average is designed to mitigate the fact that student performance can vary widely from year to year due to factors beyond a school's control such as changes in the demographic composition of student populations ("Raising The Bar," 2002). While the primary purpose of using this rolling average option is to make the school AYP measure more reliable, it can also help improve the fairness of school accountability system by reducing the chance that small schools or small subgroups within schools would be left out of reporting due to the states' minimum group size (N) requirement. Moreover, it needs to be noted that the uniform averaging option also has some potential to help

struggling schools meet the AYP target under the circumstance of declining test scores. Does this option really work to save a school with downward performance trend from being identified by the state as failing AYP?

Second, the safe harbor provision is designed to address a fairness issue: Does AYP measure school progress in a way that different groups of students in the same school can meet the same performance target at different rates? Basically, the law requires that schools disaggregate the test results into subgroups (e.g., major racial/ethnic groups, economically disadvantaged students, students with disabilities, English Language Learners) and have all of them meet the same AYP target. This requirement has the danger of assuming that all categories will move forward at the same rates (NECEPL, 2002). However, the NCLB also gives schools the option of a "safe harbor", which is designed to lesson the difficulty of reaching the same AYP target for all groups of students at the same rates and give academically viable schools a second chance. For school where the performance of one or more student subgroups on one or both of reading and math assessments fails to meet AYP targets, the school will be considered to have reached AYP under this provision if the percentage of students in that group who failed to reach proficiency decreased by 10 percent from the preceding year and also the group made progress on another academic indicator. Is this option powerful enough to save an at-risk school from being identified by the state as failing AYP?

It was estimated that up to 80 percent of schools in some states could be targeted as needing improvement or corrective action in the first few years (Marion et al., 2002; Olson, 2002, April 18). These earlier predictions from state simulations used only student assessment results without looking at test participation rates, other academic indicators, or "safe harbor" provisions under the NCLB (Marion et al., 2002). Since those earlier predictions came before the U.S. Department of Education's guidance or regulations for AYP, it was pointed out that some of the interpretations states have used in building their projections may not have taken advantage of all the options available (Olson, 2002, April 18). Therefore, we need new predictions with the options enabled, and the result may or may not differ from the earlier predictions.

In this paper, I focused on the issue of feasibility and investigate several "what if" questions through simulation analyses of the data collected from Maine and Kentucky schools during the 1990s: how the NCLB's AYP formula would have worked if we had applied it to past school performance data and what would have happened if we had applied options that the current formula permits. Specifically, the objective of this study was to (a) investigate the feasibility of the current AYP requirements for schools and (b) explore the impact of using "uniform averaging (rolling average)" and "safe harbor" options on the AYP results. I examined whether and how application of "rolling average" and "safe harbor" provisions improve the chance of meeting AYP target over the long run and at the same time reduce the risk of failing to meet the AYP for 2-5 consecutive years. The answer to questions of who might win or lose from the current AYP race and how we can make this measurement-driven accountability strategy more realistic and fair for all may provide insight that will guide policymaking.

Data and Methods

Aggregate school performance data from all public schools in two states, Kentucky and Maine, were collected and examined. Early on, both states (a) established student assessment systems to monitor their schools' academic progress and (b) made a greater effort to align their assessments with their content and performance standards (Lee & McIntire, 2002). Despite these common characteristics, the two states' assessments differed significantly in terms of the stakes attached to the assessment results: high-stakes test in Kentucky vs. low-stakes test in Maine. The 8th grade mathematics achievement data collected from the two states' student assessments were used for analysis: the Kentucky Instructional Results Information System (KIRIS) for the 1993-98 period and the Maine Educational Assessment (MEA) for the 1995-98 period. Because both states changed their state assessments since 1999, and the results were not directly comparable to old ones, all of these analyses were restricted to the pre-1999 period. Using only data collected after the NCLB legislation was not considered to be a viable option, because the data were available for only one or two years and they were not sufficient for an estimation of the longer-term consequences.

In congruence with the NCLB AYP requirements, standards-based interpretation of the test results were applied to determine academic performance of students against the performance standards set by the state. For the Maine data, the percentage of students scoring at or above "Advanced" level on the 1995-1998 MEA was used; for Kentucky, the percentage of students scoring at or above "Proficient" on the 1993-1998 KIRIS was used. Both "Advanced" and "Proficient" levels were next to the highest among four achievement levels and can be regarded as meeting state performance standards. Indeed, these two states' proficiency standards were set at a highly comparable level (in Kentucky) or at an even higher level (in Maine) than their corresponding proficiency standard on the National Assessment of Educational Progress (NAEP). For example, the percentages of 8th grade students in Kentucky who turned out to perform at or above Proficient level in mathematics as of 1996 were 16 on the NAEP and 14 on the KIRIS; the corresponding percentages in Maine were 31 on the NAEP and 9 on the MEA.

First of all, the current AYP rules were used to determine baseline and annual AYP targets in each state: the percentage of students proficient in a school at each state's 20th percentile rank in the first available year was used as the baseline AYP target. On top of that baseline, equal increments were made every year so that the AYP target becomes 100 in 12 years. Therefore, the baseline AYP target for Maine schools was set to be zero in 1995, and the subsequent AYP target added an increment of 8.3 every year to make its ultimate target equal to 100. Likewise, the baseline AYP target for Kentucky was set to be 8.8 in 1993, and the subsequent AYP target added an increment of 7.6 every year to reach 100 in 12 years from the baseline.

Given such hypothetical AYP target lines, Figure 1 and Figure 2 show the distributions of school AYP measures (i.e., the percentage of 8th grade students deemed proficient on the state math assessment) respectively in Maine and

Kentucky. In Maine, schools made very modest amount of gain, that is, about 1 percent gain per year on average so that they got farther and farther behind the AYP target over time (see Figure 1). In 1996 (Year 2), more than half of the schools in Maine were already performing below the AYP target, and a large majority of schools were so two years later (Year 4). While schools in Kentucky made relatively larger achievement gains (on average 3 percent gain per year) than their counterparts in Maine during the period, they also could not have caught up with the AYP target that grew more rapidly (see Figure 2).

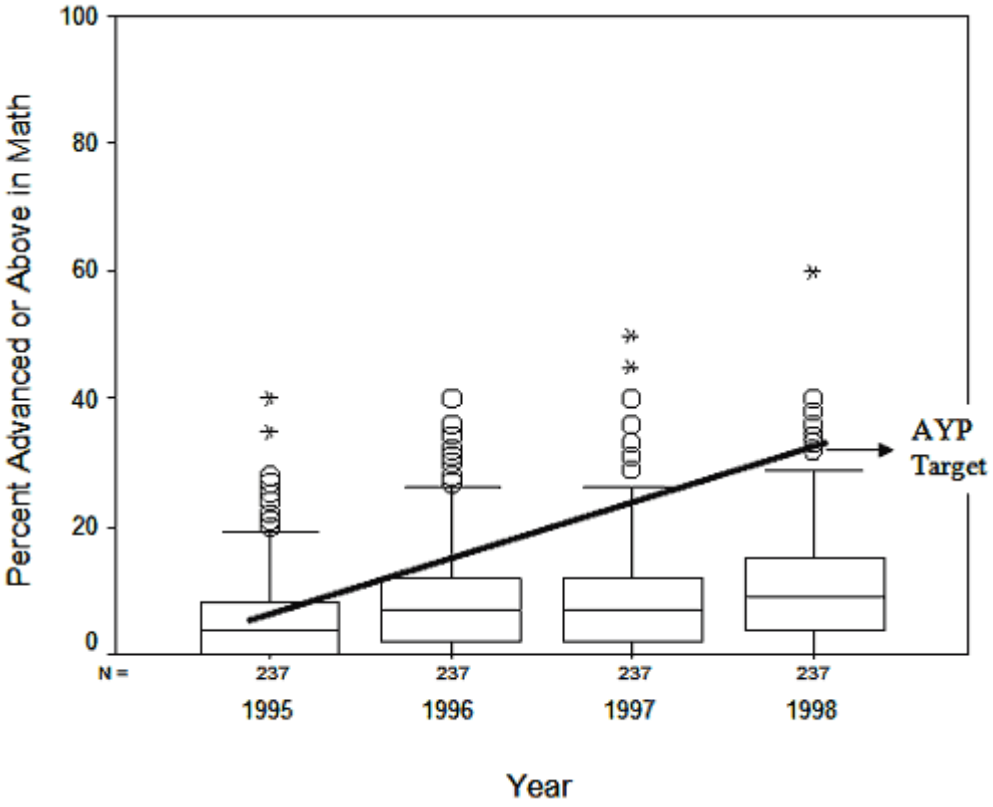

Figure 1. Maine Schools' 1995-98 Performance Trends against Hypothetical AYP Targets in 8th Grade MEA Mathematics

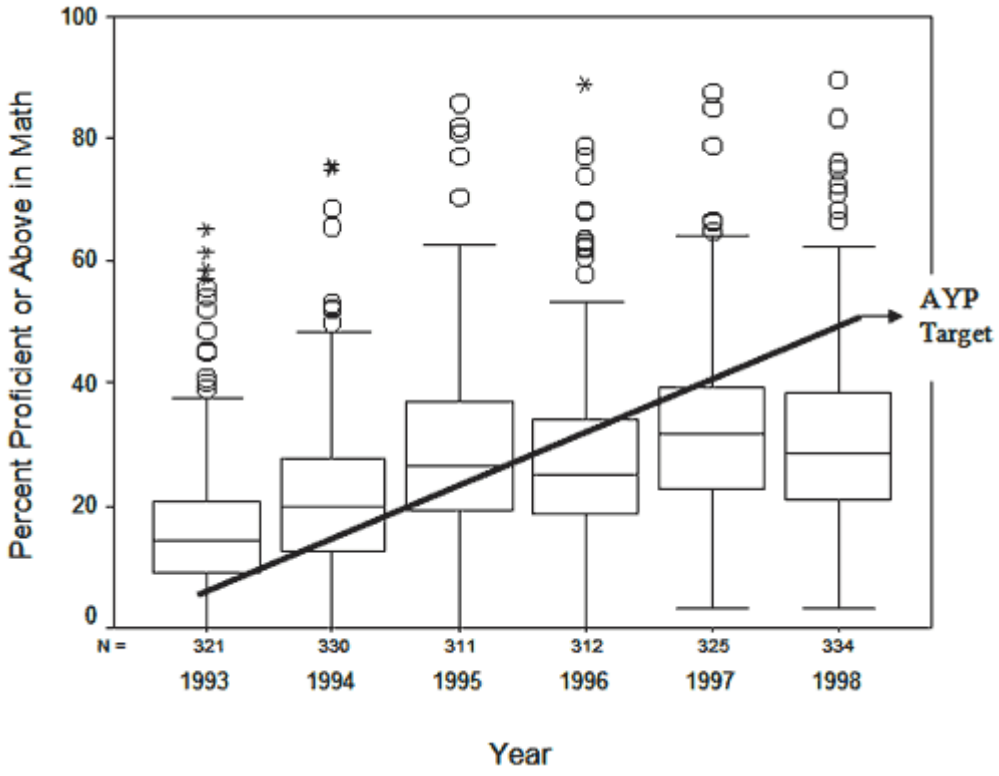

Figure 2. Kentucky Schools' 1993-98 Performance Trends against Hypothetical AYP Targets in 8th Grade KIRIS Mathematics

Assessing the effect of the "Rolling Average" option on school AYP

Under the "rolling average" (uniform averaging procedure) provision, it is assumed that schools can average test scores from the current school year with test scores from the preceding one or two years. This works in a school's favor when test scores decline but it works against a school when scores rise. If this rolling average option is used every time regardless of individual schools' variable growth patterns, it can result in a greater number of schools being identified as failing to meet AYP every year. This could have happened in both Maine and Kentucky because their schools on average made progress over the course of 4 or 6-year periods. In this study, it was assumed that the rolling average procedure was used by schools only when they obviously benefited from the option (i.e., when school performance declined). According to Scott Marion, who was the co-chair of the Joint Study Group on Adequate Yearly Progress (AYP) and co-authored a report (Marion et al., 2002), this assumption may not be unreasonable: "To be fair, schools shouldn't be able pick and choose when they can use the multi-year average. However, we've suggested that the state set up an appeal process whereby schools that miss AYP because of the earlier years included in the multi-year average be granted an appeal. So it is sort of like picking and choosing when to apply multi-year averages, but it occurs through the appeal process." (Personal communication, March 18, 2004). Nevertheless, whether states would actually allow schools to use the rolling average option in such a flexible way remains an open question

(see Erpenbarch, Forte-Fast, Potts, 2003 for examples of state plans).

The following rule was employed in this simulation's determination of using rolling average for AYP calculation: If the rolling average score (i.e., the mean of scores from current year plus preceding two years) is greater than current year score, then the rolling average is used; otherwise the current year score is used instead. Simple averaging method was used without any weighting.

If $(X_{t-2} + X_{t-1} + X_t)/3 > Xt$, then $AYP = (X_{t-2} + X_{t-1} + X_t)/3$

Otherwise $AYP = X_t$

where X_{t-2} = Percent proficient at year t-2, X_{t-1} = Percent proficient at year t-1, X_t = Percent proficient at year t (current year)

Simulation analyses of the estimates of schools that would have failed to meet the AYP target with or without this rolling average procedure were conducted. Because sanctions may apply to schools which fail to meet AYP for two or more consecutive years, the focus of this analysis was schools that belong to this high-risk category. Some schools which may fail often but not in a row would not be designated as "in need of improvement" according to the regulation. Odds ratio was computed to compare the relative risk of failure with vs. without using the rolling average option.

Assessing the effect of the "Safe Harbor" option on school AYP

The "safe harbor" provision applies to schools in which one or more of the subgroups of students fail to reach their uniform, schoolwide AYP target. According to the provision, the school shall be considered to have made adequate yearly progress if the percentage of students in that group who did not meet or exceed the proficient level of achievement on the state assessments for that year decreased by 10 percent of that percentages from the preceding school year and that group made progress on one or more of academic indicators. Although this option implies giving some recognition to schools which have made certain minimum level of progress for every subgroup despite its uneven success among different subgroups, the amount of progress required for this safe harbor application varies among subgroups; the school has to demonstrate a greater progress for a subgroup which performs at a relatively lower level in terms of its percent proficient students. While the uniform averaging procedure can also be used to combine multiple years' data for the safe harbor review, there are variations among different states in their approaches to addressing the inherent instability of gain scores (see Erpenbarch, Forte-Fast, Potts, 2003 for examples of state plans).

To examine how the "Safe Harbor" option would work for low-income students, one of the subgroups as identified by students who were eligible for free or reduced-price lunch, was chosen. Before the NCLB legislation, disaggregated student performance data was hardly available. The school aggregate performance data collected from both Maine and Kentucky was not an exception to this conventional reporting pattern as they did not break down the aggregated results by demographic subgroups. In the absence of school-level

data on the achievement of students in free/reduced school lunch program, the statewide average achievement results based on the NAEP 1996 8th grade state math assessment were used for estimation. At the same time, the absence of data on another academic indicator (e.g., performance on another type of test or retention/promotion rate) precluded an application of the requirement.

The percent students at or above the NAEP proficient level was 23 for non-eligible students and 4 for eligible ones in Kentucky. Likewise, the percent students at or above the NAEP proficient level was 35 for non-eligible students and 18 for eligible ones in Maine. For the sake of simplifying calculations, the 21-point difference was assumed to be uniform across all schools and constant over time in Maine (see equation 1.1 below); In case of Kentucky, 21 in equation 1.1 was replaced by 17. In addition, the entire school AYP measure was specified as a function of summing each subgroup's rolling-AYP measure weighted by the percentage of students in each category (see equation 1.2 below). The following simultaneous equations were solved together to estimate each school's percent proficient free/reduced lunch students:

$$Xi - Yi = 21 \qquad (1.1)$$

$$(Xi\ Pxi + Yi\ Pyi)/100 = Zi \qquad (1.2)$$

where Xi = percent proficient students among those who are not eligible for free/reduced lunch in school i; Yi = percent proficient students among those who are eligible for free/reduced lunch in school i; Zi = percent proficient students total in school i; Pxi = percent students who are not eligible for free/reduced lunch in school i; Pyi = percent students who are eligible for free/reduced lunch in school i (i.e., 100 – Pxi).

In the above equations, Zi, Pxi, and Pyi are known variables available from the data and their values are used to estimate Xi and Yi. With the estimated percentage of free/reduced lunch students who are proficient in each school at year t (Y_t), the following safe harbor rule was applied to schools which otherwise would fail to meet the AYP target for free/reduced lunch students: If $((100 - Y_t) - (100 - Y_{t-1})) \geq (100 - Y_t)/10$, then schools would be regarded as meeting the AYP target for free/reduced lunch students. It was assumed that the group made progress on another academic indicator. Odds ratio was computed to compare the relative risk of failure with vs. without using the safe harbor option.

Results

When using the current AYP goal and timeline (100% proficient within 12 years) on retrospective school performance data (1993-98 in Kentucky and 1995-98 in Maine), the percentage of schools that would meet their AYP target overall turned out to decrease exponentially over the course of the first few years (see Table 1). In Kentucky, it was 80 percent in the first year, plummeted to 36 percent in the 4th year, and further down to 10 percent in the 6th year. In Maine, it started as 100 percent in the first year (because baseline AYP goal was set to 0), became 44 percent in the 2nd year, and dropped down to 6 percent in the

4th year. This implies that most schools would have enormous difficulty meeting the NCLB AYP requirement that appears to be an unrealistic expectation given a relatively high performance standard (proficient) and a relatively short time line (12 years).

Even when the rolling average option was used, it would have only slightly increased the chance of schools' meeting the AYP target (see Table 1). The odds of meeting AYP target with the rolling average was only 1.06 - 1.24 times greater than the odds of meeting AYP target without the rolling average. With the rolling averaging option, the percentage of schools that would meet their AYP target in the 2nd year, for example, may increase from 44.3 to 46 in Maine and from 35.9 to 39.5 in Kentucky. This implies that the rolling average has very weak potential to save schools from being identified as failing when their scores decline.

Table 1. Percentage of Maine and Kentucky Schools that would Meet AYP Target with vs. without Rolling Average Option

	Maine			Kentucky		
Year		Rolling	OR	No Rolling	Rolling	OR
1	100.0			80.4		
2	44.3	46.0	1.07	66.1	69.5	1.17
3	10.5	12.7	1.24	61.4	62.8	1.06
4	5.9	7.2	1.24	35.9	39.5	1.17
5				28.3	30.5	1.11
6				10.3	10.9	1.07

Note: OR is the odds ratio of given percentages, i.e., the ratio of the odds of schools meeting the AYP target for all students each year with a rolling average of their corresponding odds of passing without the rolling average option.

The percentage of schools that would fail to meet AYP for two consecutive years at least once was very high: 75 percent in Kentucky and 87 percent in Maine (see Table 2). While the risk tends to drop significantly for the longer periods, it still remains a substantial threat to most schools. The failure rate for three years in a row would be as high as 57 percent in Kentucky and 52 percent in Maine. Although the failure rate for 5 consecutive years was less than 10 percent in Kentucky for the 6-year period, the risk would have been much greater for full 12-year cycle.

Table 2. Percentage of Maine and Kentucky Schools that would Fail to Meet AYP Target for 2-5 Consecutive Years with vs. without Rolling Average Option

	Maine		Kentucky		
Frequency	Rolling	OR	No Rolling	Rolling	OR

2 Years	87.3	86.5	.93	75.2	73.3	.91
3 Years	51.9	50.2	.93	57.1	55.9	.95
4 Years	0.0	0.0		17.7	17.1	.96
5 Year				8.5	8.8	1.04

Note: OR is the odds ratio of given percentages, i.e., the ratio of the odds of schools failing to meet the AYP target for free/reduced lunch students for 2-5 years in a row with safe harbor to their corresponding odds of consecutive failure without the safe harbor option.

The use of the rolling average procedure helps reduce consecutive failure rates in both states. As with the single-time failure rate, however, the degree of this risk reduction tends to be very small (see Table 2). The odds of failing to meet AYP target for consecutive years with the rolling average is .91 - 1.04 times greater than the odds of failing without the rolling average.

Applying the AYP target to a subgroup of low-income students (i.e., students who receive free/reduced lunch in this analysis) increases the risk of school failure about two to three times. The percentage of schools that would meet the AYP target for this particular disadvantaged group in Year 2 is only 6 in Maine and 32 in Kentucky (see Table 3). These figures were much smaller than corresponding figures estimated with the entire group of students in each school (cf. Table 1).

Table 3. Percentage of Maine and Kentucky Schools that would Meet AYP Target for Low-Income Students (Eligible for Free/Reduced Lunch) with vs. without Safe Harbor Option

	Maine			Kentucky		
Year	No Safe Harbor	Safe Harbor	OR	No Safe Harbor	Safe Harbor	OR
1	100.0			36.7		
2	5.7	7.2	1.28	31.9	36.4	1.22
3	1.4	5.3	3.94	30.8	37.2	1.33
4	0.5	2.4	4.89	13.9	33.8	3.16
5				10.9	20.2	2.07
6				3.4	29.1	11.66

Note: OR is the odds ratio of given percentages, i.e., the ratio of the odds of schools' meeting the AYP target for all students each year with safe harbor to their corresponding odds of passing without safe harbor option.

Using the "safe harbor" option increases the chance that schools would meet the AYP target for free/reduced lunch students (see Table 3). The odds ratio for meeting AYP target with the safe harbor ranges from 1.22 to 11.66. However, this might have overestimated the effect because the requirement of making progress on another academic indicator was not considered. At the same time,

using the safe harbor option reduces the risk of being identified as a failing school for consecutive years and facing undesirable consequences (see Table 4). The odds ratio for failing to meet AYP target for 2-5 years in a row with the safe harbor ranges from .23 to .75. Even with this option, however, the risk remains high, and up to 90 percent of schools will be regarded as needing improvement. While this estimation was based on only one subgroup, that is, economically disadvantaged students, simultaneous evaluation of other subgroups including students with learning disabilities and LEP/ELL students may result in greater failure rates.

Table 4. Percentage of Maine and Kentucky Schools that would Fail to Meet AYP Target for Low-Income Students (Eligible for Free/Reduced Lunch) for 2-5 Consecutive Years with vs. without Safe Harbor Option

	Maine			Kentucky		
Year	No Safe Harbor	Safe Harbor	OR	No Safe Harbor	Safe Harbor	OR
2 Years	98.6	94.3	.23	94.1	86.9	.42
3 Years	93.8	91.9	.75	84.1	66.8	.38
4 Years	0.0	0.0		49.2	39.4	.67
5 Year				37.0	39.5	.71

Note: OR is the odds ratio of given percentages, i.e., the ratio of the odds of schools' failing to meet the AYP target for free/reduced lunch students for 2-5 years in a row with safe harbor to their corresponding odds of consecutive failure without safe harbor option.

Now we can compare all the results of this simulation analysis under four different scenarios: (1) applying AYP to the entire group of students schoolwide without using the rolling average and safe harbor options, (2) applying AYP to the entire group of students schoolwide with the rolling average option only, (3) applying AYP to the entire group of students schoolwide as well as the subgroup of free/reduced lunch students with the rolling average option but without the safe harbor option, and (4) applying AYP to the entire group of students schoolwide as well as the subgroup of free/reduced lunch students with both the rolling average and the safe harbor options. The results that would be obtained under the above-mentioned four different scenarios are compared in Figure 3 and Figure 4 with abbreviated labels of each scenario: (1) No Rolling Average, (2) Rolling Average, (3) No Safe Harbor, and (4) Safe Harbor.

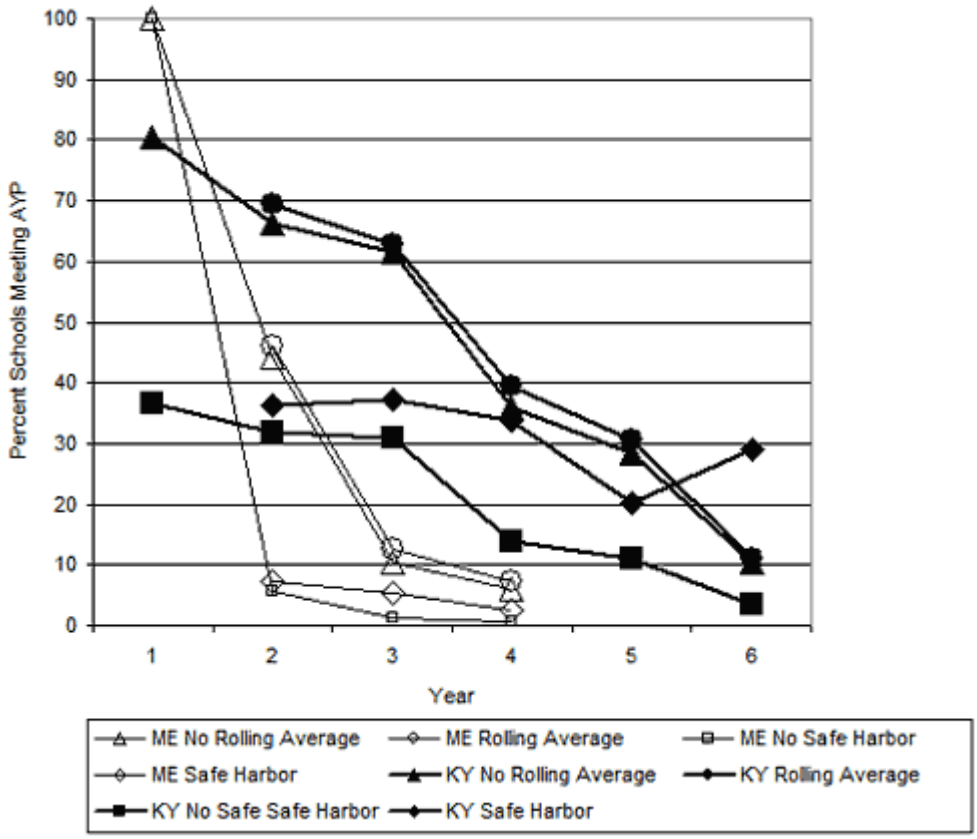

Figure 3. Percentages of schools in Maine and Kentucky that would meet AYP Target under different options.

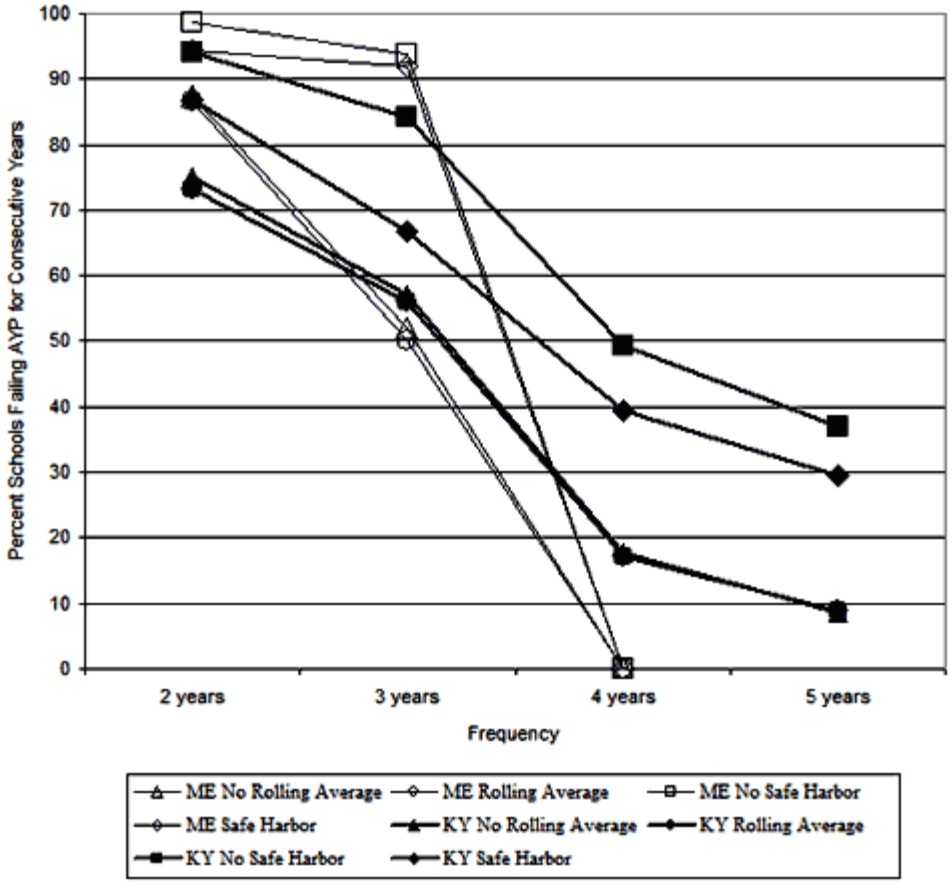

Figure 4. Percentages of schools in Maine and Kentucky that would fail to meet AYP for 2-5 years in a row.

First of all, we apply the AYP target to the entire body of students but not to subgroups in each school and do not use the rolling average and safe harbor options (see “No Rolling Average” lines in Figure 3 and Figure 4). Such schoolwide application of the AYP formula without looking into subgroups was what the states typically did for evaluating school AYP before the NCLB legislation. By using the rolling average option schoolwide, we can show some improvement in the chance of schools meeting AYP each year and for consecutive years as well, but the difference is highly marginal (see “Rolling Average” lines in Figure 3 and Figure 4). Now by applying AYP to a group of low-income students as the NCLB requires, we see substantial increases in the risk of school failure (see “No Safe Harbor” lines in Figure 3 and Figure 4). By and large, the comparison shows the benefit of using the safe harbor option, but it also reveals that the option is not strong enough to save many struggling disadvantaged schools from the risk (see “Safe Harbor” lines in Figure 3 and Figure 4).

Discussion

Policy implications of this study need to be discussed carefully given the fact that the findings are based on the simulation analysis of the past school performance data in a single grade and a single subject area from two selected

states. It needs to be noted that the study has some unwarranted assumptions about school AYP measures and targets within the parameters of the NCLB and that the actual results can be quite different if the two states make different choices (e.g., using an index measure of AYP, increasing the AYP target in a nonlinear, stepwise fashion). Whatever estimation methods used, this study might underestimate or overestimate the schools' future progress expected under this new legislation, NCLB. The results may have been different if schools had faced in the past the stronger incentives embodied in current AYP rules. Moreover, the results might be different if the performance standard used in the past is significantly higher or lower than the current performance standard adopted under new testing systems in both states. However, the comparison of Kentucky and Maine (high-stakes testing vs. low-stakes testing environments with their commonly challenging state assessments and high performance standards) can give us an insight into possible consequences of the NCLB AYP policy for schools across the nation.

With these caveats in mind, the results of this simulation analysis turn out to provide very gloomy projections of schools' chance to meet the AYP target, warning federal and state education policymakers against massive school failure under the NCLB. It does not appear to be feasible for many schools across the nation to meet the current AYP target within its given 12-year timeline. It is not realistic to expect schools to make unreasonably large achievement gains compared with what they did in the past. Many schools are doomed to fail unless drastic actions are taken to modify the course of the NCLB AYP policy or slow its pace. Contrary to some expectations, using both rolling average and safe harbor options does not work to reduce the risk of massive school failure. Although the rolling average can help improve more stable estimation of school performance, it hardly reduces the risk of school failure. The safe harbor option also fails to provide a strong safety net to at-risk schools despite what its name implies.

When a majority of schools fail, there will not be enough model sites for benchmarking nor enough resources for capacity building and interventions. This situation can raise a challenging question to the policymakers: is it school or policy that is really failing? There is a potential threat to the validity of the NCLB school accountability policy ultimately if such prevailing school failure occurs as an artifact of policy mandates with unrealistically high expectations that were not based on scientific research and empirical evidence.

One approach that policymakers can consider to make the AYP targets more realistic and fair might be to use an effect size measure for guidance. For example, one might reasonably expect that schools should make progress every year by say 20% of the standard deviation of school-level percent proficient measure; this amounts to about 2.5 - 3.0 percent in Kentucky and 1.5 – 2.0 percent in Maine. This amount of progress may be regarded as small by conventional statistical standard (Cohen, 1977), but it is exactly what an average school in both states managed to accomplish in the past. In a similar vein, one can consider setting the safe harbor threshold for a subgroup at certain percentage of the standard deviation (e.g., reduce the percentage of non-proficient low-income students by 10% of the standard deviation). A similar suggestion along with the use of scale score rather than percent proficient was

made by other analysts (Linn, Baker, & Betebenner, 2002).

While using an effect size metric with scale scores may help set more realistic performance targets and better recognize schools' academic progress, it is not permissible under the current law. This idea also raises questions as to whether to use standard deviation of student-level test scores or school-level average test scores and whether to derive the standard deviation from original test score variance or residual variance with adjustments for demographic differences among students and their schools. In Maine and Kentucky, the school-level standard deviation was only 40 percent of the student-level standard deviation of mathematics achievement scores. Once the differences among schools in their students' racial and socioeconomic background characteristics, the adjusted school-level variance of residuals is reduced further down to the half of original school-level variance (see Lee & Coladarci, 2002 for the analysis of within-school vs. between-school math achievement distributions in Maine and Kentucky).

Using different methods with different measures would produce different results and, consequently, different conclusions. Whether one prefers a criterion-referenced or norm-referenced approach to setting AYP target and evaluating school progress, the ultimate concern is not simply improving the feasibility of schools' meeting their AYP targets in the short term but rather enhancing the schools' capacity for sustained academic improvement over the long haul. Given limited amount of resources available from the federal government and limited capacity of the state agencies as well, reducing the identification of schools in need of improvement would help states provide more targeted assistance to a smaller number of disadvantaged schools which have a large number of at-risk students. Nevertheless, applying the AYP options such as rolling averages and safe harbor had better not be compromised by future prospect of limited support and short-term interests in reducing school identifications. The long-term success of school accountability system does not depend on the number of passing schools but on the results of student achievement.

Note

This article is based upon work supported in part by the National Science Foundation under Grant No. 9970853. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This study simply utilizes the past school performance data from Maine and Kentucky for simulation analyses, but all assumptions, results, and interpretations given in the article have nothing to do with the two states' current AYP policies and outcomes. An earlier version of this paper was presented at the 2003 AERA annual meeting in Chicago. E-mail JL224@buffalo.edu for correspondence about this manuscript.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Erpenbarch, W. J., Forte-Fast, E., Potts, A. (2003). *Statewide Educational Accountability Under NCLB: Central Issues Arising from an Examination of State Accountability Workbooks and U.S. Department of Education Reviews Under the No Child Left Behind Act of 2001.* Washington, DC: Council of Chief State School Officers. Available at http://www.ccsso.org.

Hill, R. (1997). Calculating and reducing errors associated with the evaluation of adequate yearly progress. Paper presented at the annual assessment conference of the Council of Chief State School Officers (ERIC Publication No. ED 414307).

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores. In D. Ravitch (Ed.). *Brookings Papers on Education Policy 2002*. (pp. 235-284). Washington, DC: Brookings Institution.

Kim, J., & Sunderman, G. L. (2004). *Large Mandates and Limited Resources: State Response to the No Child Left Behind Act and Implications for Accountability*. Cambridge, MA: The Civil Rights Project at Harvard

La Marca, P. M. (2003). Factors affecting the statewide implementation of an adequate yearly progress model. Paper presented at the annual meeting of the American Educational Research Association in Chicago.

Lee, J. (2003). Evaluating Rural Progress in Mathematics Achievement: Threats to the Validity of "Adequate Yearly Progress." *Journal of Research in Rural Education, 18*(2), 67-77.

Lee, J. & Coladarci, T. (2002). *Using Multiple Measures to Evaluate the Performance of Students and Schools: Learning from the Cases of Kentucky and Maine*. Orono, ME: University of Maine. Available at http://www.ume.maine.edu/naep/SSI

Lee, J, & McIntire, W. (2002). *Using National and State Assessments to Evaluate the Performance of State Education Systems: Learning from the Cases of Kentucky and Maine*. Orono, ME: University of Maine. Available at http://www.ume.maine.edu/naep/SSI

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*, 3-16.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32(7), 3-13.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24(1), 29-36.

Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., Sheinker, J. (2002). *Making valid and reliable decisions in the determination of adequate yearly progress.* A Paper in the Series: Implementing The State Accountability System Requirements Under the No Child Left Behind Act of 2001. Washington, DC: Council of Chief State School Officers. Available at http://www.ccsso.org.

New England Center for Educational Policy and Leadership (2002). *Implementing the No Child Left Behind Act of 2001: A tool kit for New England state policy makers.* Storrs, CT: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107-110.

Olson, L. (2002, April 18). 'Inadequate' yearly gains are predicted. *Education Week*. Available at http://www.edweek.org/ew/ew_printstory.cfm?slug=29ayp.h21.

Raising the bar: The complexities of "adequate yearly progress." (2002) *Education Assessment Insider, 1(5),* 5.

Thum, Y. (2002). Design of School Performance and School Productivity Indicators: Measuring Student and School Progress with the California API. Working draft.

About the Author

Jaekyung Lee

Graduate School of Education
SUNY at Buffalo
E-mail: JL224@buffalo.edu

Jaekyung Lee is an assistant professor of education at University at Buffalo, the State University of New York. He was National Academy of Education/Spencer Postdoctoral Fellow and Principal Investigator of NSF Statewide Systemic Initiatives (SSI) study. His current research focuses on the issues of educational accountability and equity.

The World Wide Web address for the *Education Policy Analysis Archives* is
epaa.asu.edu

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu.

EPAA Editorial Board

EPAA is published by the Education Policy Studies Laboratory, Arizona State University