



A Union of Professionals

WHAT'S WRONG WITH THE NO CHILD LEFT BEHIND ACT'S (NCLB) ADEQUATE YEARLY PROGRESS (AYP) FORMULA? A Summary of Internal and External Research Findings¹

Although the P in AYP stands for “progress,” AYP does not measure progress.²

- AYP does not measure the *same* students over time (e.g., from one grade to the next or even at the beginning and then at the end of the school year), so it is not a *progress* measure at all. Instead, it measures the achievement *status* of *different groups* (cohorts) of students at one point in time in any given year (e.g., the percentage of last year’s fourth-graders who hit at least the proficient cut score on the state’s tests, the percentage of this year’s fourth-graders who did so, next year’s percentage, and so on). AYP does not even tell you where these different groups started but only where they ended, and then only in terms of the percentage that made it to the proficient level. (See the next section for how “progress” is measured in Safe Harbor.)
- The P in AYP only means the progressively higher state AYP targets a school has to meet, regardless of whether any given cohort of its students started well above or well below the annual, or even the starting, state AYP targets.
- Because AYP does not measure progress, it cannot discern whether or not a school has the requisite annual percentage of proficient students because of where its students started—an invalid basis for accountability—or because of its effectiveness in improving achievement, the only valid and fair basis for accountability.

Although Safe Harbor is touted as crediting progress and therefore being the kinder, gentler passage to AYP,³ it is at least as tough as regular AYP, it rarely works, and it even more rarely produces statistically reliable decisions.⁴

- Safe Harbor, like regular AYP, does not measure the *same* students over time. It compares the percentage of proficient-level students in, say, last year’s 10th grade to the percentage in this year’s 10th grade—again, *different* groups of students. If the decrease in the percentage of non-proficient students in the two different cohorts is at least 10 percent, for the school and for each subgroup failing regular AYP, the school reaches Safe Harbor.
- The Safe Harbor measure (change scores on tests from different cohorts of students) is notorious for its statistical unreliability. The smaller the number of students Safe Harbor measures, the greater the statistical unreliability of Safe Harbor decisions—and the greater the odds that

¹ This brief is based primarily on AFT research, using state databases, on how AYP worked in Maryland, Massachusetts, New Hampshire, New York (in progress), Ohio and Pennsylvania in the first year of NCLB implementation. It also uses external research on AYP in California, Colorado, Connecticut, Florida, Louisiana, Minnesota, North Carolina and Washington, and a multistate analysis of student achievement growth. The focus is on AYP’s academic criteria. Factoring in test-participation rate and attendance/graduation rate criteria only multiplies and worsens the problems this research identifies.

² ASR-CAS (2002); Linn (2001); Hamilton et al.

³ See, for example, Hall et al.; The Education Trust; Progressive Policy Institute.

⁴ Hill & DePascale (2003a, b); Kane & Staiger (2002, 2003); Linn & Haug; ASR-CAS; Lee; Nelson & Rosenberg (2004).

subgroups and schools will be misclassified one way or the other. The greater the number of students measured, the lower the odds of getting a statistical lucky break and of reaching AYP through Safe Harbor.

- The Safe Harbor growth targets must be met in each subject for each AYP-counted subgroup of students that failed to reach AYP's regular proficiency targets. As is the case with regular AYP, schools with many counted subgroups often have to meet several Safe Harbor growth targets—a large educational and statistical challenge. In Massachusetts, for example, 93 subgroups met Safe Harbor growth targets in reading and 42 subgroups did so in math, but only 44 schools made AYP through Safe Harbor. In New Hampshire, only 12 schools made AYP through Safe Harbor. The 141 schools that still failed AYP required 329 separate Safe Harbor calculations, only 31 of which met the Safe Harbor improvement targets.
- Not surprisingly, then—and as even the Education Trust report that defends AYP on the basis of Safe Harbor inadvertently shows—Safe Harbor schools tend to be relatively small, with few or no subgroups large enough to count for AYP. Safe Harbor schools also include a disproportionate number of magnet and other selective schools. It is likely, then, that the different cohorts of students being measured were already near or at proficient relative to schools that did not reach Safe Harbor. In Massachusetts, for example, Safe Harbor helped smaller-than-average schools with high average scores and few subgroups. Pennsylvania schools making AYP through Safe Harbor averaged only 53 tested students, compared to 141 in the typical school, and, not surprisingly, had fewer subgroups.
- Predictably, few schools reach Safe Harbor even when they make a lot of progress. In Massachusetts, it was 44 schools (2.7 percent of schools); in New Hampshire, 12 schools (2.7 percent of schools); and in Pennsylvania, it was 18 schools (1.0 percent of schools).⁵

Although NCLB aims to hold *all* schools accountable, AYP only does so selectively. Making or failing AYP (and Safe Harbor) is strongly influenced by a number of non-academic factors, such as a school's size, the number of its grades that are tested by the state, and the size and number of its subgroups—not to mention the laws of statistics.⁶

- Larger schools and/or schools with more state-tested grades are more likely to have a subgroup whose size is large enough to count separately in AYP. If School A has a counted subgroup, but School B's identical subgroup is not large enough to count in AYP—and both schools have average proficiency levels that meet the state's AYP targets—School A will fail AYP if its subgroup does, while School B will not, even if its subgroup students score lower than the subgroup in School A.
- On average, special education students score very low on regular tests. Therefore, when a school has an AYP-counted special education subgroup, it almost invariably fails AYP—*even when its special education subgroup scores higher than special education students in a school where that subgroup is not large enough to count in AYP*. In Pennsylvania, schools with an AYP-counted

⁵ Nelson & Rosenberg. Some state databases make it impossible to determine Safe Harbor schools, but the evidence indicates that these paltry Safe Harbor rates are typical.

⁶ Nelson; Linn et al. (2003a, b); Kane & Staiger (2002, 2003); Hill & DePascale (2003a, b, c); Lee; Novak & Fuller; Minnesota; ASR-CAS. Nothing in this section should be construed as criticism or support of minimum N's or states' use of confidence intervals, unless otherwise noted. That vital discussion is outside the bounds of this brief.

special ed subgroup had a 97 percent AYP failure rate; in Massachusetts, the comparable figure was 65 percent; in Maryland, 56 percent; and in New Hampshire, 57 percent.⁷

- The more subgroups a school has, the lower its odds of making AYP—not only because subgroup proficiency levels, on average, were below state AYP targets to begin with, but also because it is *statistically* harder for multi-subgroup schools to make AYP than it is for schools with fewer or no AYP-counted subgroups, *even when their average achievement is the same*. In Ohio, schools with more than one subgroup were five times as likely to miss AYP targets as one-subgroup schools (the one usually being a white subgroup). Pennsylvania schools with more than one subgroup were at least three times more likely to fail AYP as one-subgroup schools. In Maryland and Massachusetts, multi-subgroup schools were four times more likely to fail.
- In some states, small schools fell outside of AYP accountability altogether. In most states in 2002-03, entire schools—typically larger ones—were judged on their “effectiveness,” including receiving sanctions, based on AYP calculations for students in only one, and sometimes two, of their grades.⁸ Under NCLB, no state may hold private schools that receive Title I funds accountable; but every state must apply AYP to *all* of its public schools, even those that do not get Title I funds. The law holds K-2 (untested grades) “feeder schools” accountable for their receiving schools’ percentage of proficient students and test-participation rate.

Although NCLB is presumed to hold all schools accountable for achievement gaps, AYP only does so selectively.

- Schools with subgroups that are not large enough to count separately in AYP are not held accountable for achievement gaps, no matter how large the gaps may be. The only schools held accountable for achievement gaps, no matter how small the gaps may be, are those with even just one AYP-counted subgroup whose percentage of proficient students falls below even just one of the state’s AYP targets in any given year.
- On average, achievement gaps in schools that had counted subgroups and that made AYP equaled or exceeded the gaps in schools that failed AYP. Nor are schools that made AYP necessarily reducing those gaps faster than those that failed AYP.⁹
- Relatively few schools with counted subgroups made AYP. These few included a disproportionate number of magnet and other selective schools, which enroll higher-achieving students in the first place; smaller schools; lower class-size schools; and schools with lower concentrations of poverty.¹⁰
- Schools that enrolled the most disadvantaged of subgroup students had the highest rates of AYP failure. Given the low starting point of disadvantaged students, it is not surprising that the large majority of a state’s subgroup students (except for the white subgroup) were enrolled in schools that failed AYP.¹¹

⁷ The minimum subgroup size for inclusion in the AYP calculation is 40 in Pennsylvania, 20 in Massachusetts, five in Maryland and 11 in New Hampshire. The latter two states use confidence intervals with margins of error that are very large for very small subgroups.

⁸ This was also the case with state and Title I accountability prior to NCLB. NCLB did not invent invalid accountability systems, but its far more stringent AYP standards and sanctions have multiplied and exacerbated the problems in such systems—and has forced states with more exemplary accountability systems to take a big step backward.

⁹ Nelson & Rosenberg; Choi et al.; Linn & Goldschmidt; Novak & Fuller; McCall et al. For additional, non-AYP-specific evidence on this, see Council of the Great City Schools (2003, 2004).

¹⁰ Nelson & Rosenberg; Novak & Fuller; The Education Trust (2001, 2002a, b); Hall et al.; Harris; Rothstein.

¹¹ Nelson & Rosenberg; Novak & Fuller; Choi et al.; Linn & Goldschmidt.

Although the A in AYP stands for “adequate,” AYP is calculated on the basis of a 100 percent proficiency goal that demands unnatural rates of “progress” from some schools and groups of students, while tolerating inadequate progress or declines among other schools and groups of students—at least for a while. Sooner or later, virtually every public school district and school will fail AYP, but not necessarily because they are “failures.”¹²

- On average, while achievement levels in disadvantaged schools are much lower than in other schools, their growth rates are the same.¹³ Therefore, schools that started substantially below even the state’s starting AYP targets—the ones enrolling the most disadvantaged children—must increase their percentage of proficient students at a rate so phenomenal that it has never been reliably evidenced. Conversely, schools that started comfortably above the targets can coast or decline and still make AYP, at least for a while. Ultimately, most schools will fail AYP.
- AYP works in a way that treats advantaged and disadvantaged schools, and selective (e.g., magnets) and non-selective schools alike, as if their students all started from the same level of achievement on the day NCLB was signed into law. AYP is also indifferent to the ironclad fact that disadvantaged youngsters, on average, are significantly behind other children before formal schooling even begins.¹⁴
- Even with an acceleration of past rates of progress, the goal of 100 percent proficiency will not be reached by 2014. Although schools enrolling large numbers of disadvantaged students, followed by large secondary schools, already fail AYP at high rates (and will likely fail many times), nearly all schools will eventually fail AYP. All but the smallest school districts will soon fail. Even though many states changed their implementation of AYP in 2003-04, these changes will only delay the inevitable failure of most schools. This is not because public education is broken or American children are deficient; it is because AYP is unrealistic, does not understand student achievement growth patterns, and does not recognize the laws of statistics.

Although the goal of 100 percent proficiency applies nationally, the states’ widely (and wildly) divergent AYP failure rates tell you nothing about the relative achievement or effectiveness of their respective public schools.¹⁵

- In the first year of AYP implementation, states’ AYP failure rates ranged from 7 percent to 85 percent. This variability is not related to differences in educational quality among the states. It is explained in some part by differences in the difficulty of states’ academic content standards; differences in the type and difficulty of the tests they use; and by differences in where states set the cut point for “proficient” on their tests. It is explained in even larger part by differences in states’ subgroups; differences in school size; differences in the number of grades they test; differences in the number they set for the size a subgroup has to reach in order to count separately in AYP (“minimum N”); whether or not the state uses “confidence intervals” in its AYP calculations; and other factors having nothing to do with educational quality. These points are reinforced by external evidence showing that there is no discernible relationship between states’ AYP failure rates and their performance and growth on NAEP.¹⁶

¹² Linn (2003a, b); Nelson & Rosenberg; Minnesota; Kane & Staiger (2003a, b).

¹³ Phillips; Jencks & Phillips; Barton & Coley; Coley (2003); Linn (2003c, d). See also Entwisle & Alexander; Council of the Great City Schools.

¹⁴ West & Germino-Hausken; Denton et al. (2000, 2001); Lee & Burkam; Hart & Risley; Barton (2003); Coley (2002); Rothstein.

¹⁵ Linn (2003a, b, c, d, e, f); Kingsbury et al.; Nelson & Rosenberg; McLaughlin & DeMello.

¹⁶ Council for Education Policy, Research and Improvement (2003).

Although NCLB stands for “No Child Left Behind,” AYP leaves many a child behind.

- AYP is indifferent to the achievement of subgroup students in schools with subgroups that are too small to count for AYP, except insofar as their achievement affects the school’s average. In Pennsylvania, for example, only 10 percent of schools had an AYP-counted special education subgroup, and in Ohio, less than 1 percent of elementary schools had this subgroup count in AYP. Native American and multi-racial subgroups hardly ever count in AYP.
- AYP is indifferent to the achievement of individual and subgroup students at any level or progression of performance other than the proficient level (although Colorado, Massachusetts, Minnesota, and New Hampshire have been allowed to give non-proficient students partial credit). Nor does AYP register declines or stagnations in achievement across and within performance levels, unless they are large enough to affect the schoolwide and/or subgroup percentage of proficient students relative to the annual AYP targets.¹⁷
- There is no reliable evidence that even highly effective schools can produce the huge gains that are necessary for the lowest scoring students to reach AYP’s escalating proficiency targets. In Pennsylvania, for example, special education subgroups had an average rating in math of 14.4 percent proficient compared to 54.9 percent for all students. In Maryland, special education subgroups had an average rating in math of 27 percent proficient compared to 50 percent for all students. Conversely, AYP implicitly signals that little or no effort needs to be made with students who are already proficient or above—or in subgroups that are not large enough to count for AYP—because AYP hardly registers them. Therefore, AYP implicitly sets up a perverse set of incentives for schools to concentrate only on raising the scores of students who are just below the proficient cut points on state tests; hard work with other groups is either futile or irrelevant in making AYP.

Since neither the A nor the P in AYP means what it says, it should not be surprising that whether or not a school makes AYP does not necessarily depend on its effectiveness.

- There is no discernible relationship between school or subgroup progress and making or failing AYP. It is where a school or subgroup starts out, relative to the state’s annual AYP targets, that counts.
- Average and subgroup achievement *growth* is often as great, or greater, in schools that failed AYP as it is in those that made AYP. Average and subgroup growth in a state’s AYP-failed schools also tends to meet or exceed the state’s average growth rate.¹⁸ In Massachusetts, for example, schools with black, Hispanic or limited English proficient (LEP) subgroups large enough to count separately in AYP determination failed AYP at rates exceeding 80 percent. Yet schools enrolling these subgroups showed greater improvement in the state’s composite proficiency index scores than the state average.
- Many schools that fail AYP are considerably more successful at making progress with their low-, average- and high-scoring students—*all* their students—than are schools that make AYP. Many schools that fail AYP are better at improving low-scoring students’ achievement, in particular,

¹⁷ Safe Harbor is a limited, flawed exception. Also, although some states use performance indexes to credit status growth at additional achievement levels, the law itself prohibits using a measure that would reduce the number of schools that AYP by itself would identify for improvement.

¹⁸ Nelson & Rosenberg; Novak & Fuller; California Department of Education; Choi et al.; McCall et al.; Linn & Goldschmidt.

than are schools that make AYP, even though schools failing AYP typically have much higher concentrations of poor children than schools that make AYP.¹⁹

- Schools with special education or LEP subgroups large enough to count separately in AYP almost invariably fail AYP—even when the achievement of these subgroup students, as well as their schools as a whole, equals or exceeds that of their peers in schools where these subgroups are too small to be separately counted in AYP.²⁰
- Although AYP cannot measure school effectiveness, it automatically deems schools that enroll large numbers of special education and low-achieving disadvantaged students “ineffective,” regardless of the progress they make with their students.

NCLB rightfully seeks to evaluate and hold schools accountable for their educational effectiveness. But the evidence shows that AYP is an unreliable and invalid measure of effectiveness that holds schools more accountable for where students start than for their academic progress and, in the case of the most disadvantaged schools, for performing an educational feat that has never been reliably evidenced.

¹⁹ Choi et al.; Linn & Goldschmidt; McCall et al.

²⁰ Nelson & Rosenberg.