NO CHILD LEFT BEHIND:
WHAT WE KNOW AND WHAT WE NEED TO KNOW

GLENN MALEYKO, ED.S.
Wayne State University
Department of Educational Leadership and Policy Studies

MARYTZA A. GAWLIK, PH.D.
Wayne State University
Department of Educational Leadership and Policy Studies

Drawing from empirically based and theoretical literature in the field, this review examines Adequate Yearly Progress (AYP) and the accountability provisions found in Title One of the NCLB legislation. States have the ability to statistically manipulate their AYP implementation, which may give a false impression to the public that AYP is a consistent measure of school effectiveness across the country. The first section identifies the measurement concerns with the implementation of AYP. The second section of the review deals with the benefits of AYP and the unintended consequences. The review also analyzes the complexities involved with establishing school accountability and the effective and ineffective provisions of the NCLB reform. Finally, this paper concludes with recommended areas of research for policymakers and educators alike who are interested in sustainable reform.

Introduction

The No Child Left Behind (NCLB) reform aims to hold educational agencies and states accountable for improving the quality of education for all students. It seeks to identify and transform low-performing schools that have failed to provide a high quality education to their students into successful schools. Furthermore, the accountability provisions in NCLB intend to close the achievement gap between high and low achieving students and especially the achievement gaps between minority and non-minority students along with the advantaged and disadvantaged students. The reform seeks to accomplish this goal using state assessment systems that are designed to ensure that students are meeting state academic and grade level content expectations (NCLB, 2002, section 101).

The implementation of the NCLB goals calls for high-level standards that are measurable for all students. There is no doubt that NCLB has provided for an increased focus on student populations that have traditionally performed at low levels (Borowski & Sneed, 2006; Guilfoyle, 2006; Haycock, 2006; Hess, 2006; Hess & Petrella, 2006; Kane, et al. 2002; Lewis, 2006) however, there are some faults with the Adequate Yearly Progress (AYP) as to whether or not AYP will be able to provide an accurate measurement of the goals that are stated in the Title One purpose statement of the NCLB legislation. Some of those faults include states being allowed to develop their own standards, test score proficiency levels, and statistical measurement formulas under AYP (Harris, 2007; Olson & Jacobson, 2006; Popham,
Measurement Concerns with the Use of AYP to Evaluate School Effectiveness

Statistical Issues, Manipulations of the AYP data, and Lack of Consistency across the Country

There are a number of problems associated with the usage of the current AYP formula in order to measure and evaluate school effectiveness. The first problem area concerns the manner in which the statistical manipulations are occurring with the implementation of AYP across the country. Borowski and Sneed (2006) conclude that the manner in which AYP is determined is arbitrary in nature. States have the ability to lower standards and manipulate statistical measures of AYP that may result in the lack of improvement in instruction.

Porter et al. (2005) found that states exercise flexibility in implementing achievement tests and this consequentially impacts whether or not schools or school districts make AYP. They focus on three specific areas of measurement that have a major impact on AYP. Those areas include: the line of trajectory that states establish en route to a 100% proficiency rate by the year 2014, the minimum number of students that are necessary in order for there to be a subgroup population that will count for AYP purposes, and whether or not the state uses a confidence interval along with how wide the confidence interval is in determining if schools or districts reach the proficiency targets that are required under AYP. The implementation of NCLB by the United States Department of Education (USDOE) provides states with some
Figure 1. Confidence Interval Example

Confidence Intervals

Porter et al. (2005) provide empirical data on the number of states that are not using confidence interval vs. those states that are using confidence intervals and the size of the confidence intervals that they are using. When examining the confidence interval decisions by states, they identified 2 states that use a confidence interval lower than 95 percent, 14 that use a 95 percent with a one-tailed approach, 1 uses a 98 percent trajectory with a one-tailed approach, 16 use a 99 percent two-tailed confidence interval approach, 6 are unspecified, and 11 do not use a confidence interval.

Popham (2005a) explains that it is not statistically sound to implement the use of a confidence interval for AYP calculations. The manipulation of the confidence intervals has a dramatic impact on the number of students and schools that pass proficiency tests and it results in an impact on the AYP status of schools. Figure 1 provides an illustration of the confidence interval, which is hypothetical for illustration purposes but not based on an actual state example. In Figure 1, the hypothetical cut score on the assessment is 65 percent with a confidence interval of ± 5 percent and the students who scored 65 percent or above are deemed to be proficient. The students who scored from 60 percent to 65 percent did not meet the proficiency cut score, but they fell within the confidence interval and they are deemed to be proficient due to the possibility of measurement error. When a state implements the use of the confidence interval students who have score within the confidence interval are calculated as proficient under the AYP formula.
The end result is the perception that schools are successful according to AYP when in fact it might have been the implementation of the confidence interval that allowed them to meet AYP standards. For example, starting in the 2004-05 school year, Michigan began to use a two level of standard error above and below student proficient test scores (Center on Education Policy[CEP], 2007; Michigan Department of Education[MDE], 2008a). Thus, students who fall within the confidence interval qualify as proficient even if they do not meet the specific cut score on the state assessment. The Michigan Consolidated State Application for the implementation of AYP (2005) that was approved by the USDOE with the use of the confidence interval resulted in more schools making AYP in Michigan (CEP, 2007). For example, in the year 2004-05 without the usage of a confidence interval in Michigan, 79 percent of elementary schools make AYP and 74 percent of middle schools make AYP. With the use of the confidence interval, 89 percent of elementary schools in Michigan make AYP and 82 percent of middle schools make AYP (CEP, 2007). Figure 2 illustrates that the use of a confidence interval can have an impact on the number of schools in a state that make AYP.

**Subgroup Size (N)**

Some studies (Meir, Kohn, Darling-Hammond, Sizer, & Wood, 2004; Olson
& Jacobson, 2006; Porter, et al., 2005) found that there are significant inconsistencies from state to state with the size (N) of subgroups that count for AYP purposes. For example, Olson and Jacobson (2006) found that the minimum N size for a subgroup in Louisiana is 10 in contrast to California where it is 50 and it must be an equivalent of 15 percent of the school population until the number reaches 100. The study includes data taken from the Associated Press from the 2003-04 school year enrollment figures. The study analyzed students in grades 3 through 8 and grade 10 since those are the grades that are tested each year under NCLB. The study analyzed five major subgroups that are counted under the NCLB AYP provision, which includes: White, Black, Asian, Native American, and Hispanic. They did not look at the other subgroups such as ELL students or special education students because there would have been the possibility that those students could have been counted more than once since they might qualify into more than one category.

Based on the AYP data, Olson and Jacobson (2006) found that over 1.9 million students across the United States are not being counted for AYP purposes due to the minimal subgroup requirements. This includes over 400,000 minority kids in California that are not being counted. The number of students that were not counted in the five subgroups that were examined includes fewer than 2 percent of White students, 10 percent of Black and Hispanic, 33 percent of Asian students, and 50 percent of American Indians. They also found that over 93 percent of the students are included in both an overall school population and in a racial or ethnic category.

The findings by Olson and Jacobson (2006) with the statistical manipulation of the subgroup size are consistent with the findings by Porter, et al. (2005) which concluded that different (N) size minimal subgroup calculations were creating inconsistencies from state to state in regards to the number of schools that are making AYP. Porter, et al. (2005) found that the subgroup sizes that states are using for AYP calculations range from 0 to 100 across the United States. There are nine states that use a minimal subgroup size that is greater than 42, fourteen states use a minimal subgroups size of 40 and twenty-seven states define a minimal subgroup of 34 or less. Figure 3 provides a sample of state minimal subgroup levels and it illustrates the disparities with the minimal subgroup size implementation among the states.
Figure 3. Subgroup minimal size in a sample of states.

In California the subgroup must equal at least 15 percent of the population otherwise the minimal could rise to 100.

In Florida if the subgroup can be as low as 50 students as long as they represent 10 percent of the school population.

Note. The data is available in the article by Olson and Jacobson (2006) and on the state department of education websites via their consolidated school AYP applications that were submitted to the United States Department of Education.

These studies (Olson & Jacobson, 2006; Porter et al. 2005) demonstrate that there are inconsistent measurement standards across the county, which produces misleading information to the public when it comes to AYP. This misleading information includes the success and failures of schools across the country under the AYP accountability provisions in NCLB. The research provides professionals and the public with important information and it gives the perception that statistical manipulation is being used so that students are not represented in specific subgroups. The studies uncover one of the problems with the AYP calculation and the inconsistency from state to state.

Lines of Trajectory

Past research (Popham, 2005a; Porter, et al. 2005; Wiley, et al., 2005) shows that the manipulation of the AYP line of trajectory has a major impact on the number of schools that make AYP. Porter, et al. provide an analysis of the number of states that are adhering to different measurement designs. They identify three specific forms or target lines of trajectory that states use with the evaluation of school effectiveness according to AYP. The researchers provide statistical data on the number of states that are using straight line achievement
trajectories, back end loaded trajectories, front end load trajectories, and straight line trajectories with plateaus. They identified 24 states that use the back loaded trajectory, 19 that use the straight trajectory with plateaus, 4 that use the straight trajectory, no states use the front loaded trajectory, and three states had an unidentified trajectory. This further exemplifies the inconsistent measures that are being implemented across the country when it comes to determining the AYP status of schools. The following figures provide examples of the line of trajectory that states might choose to use en route to a 100 percent proficiency level by the year 2014. Figure 4 provides an illustration the straightline trajectory, Figure 5 illustrates the straight line trajectory with plateaus, and Figure 6 illustrates the back end loaded trajectory.

**Figure 4. Straight line trajectory.**

![Figure 4](image_url)

Note. This figure is based on the description presented by Porter ct al. (2005) and Wiley ct al. (2005) which is aligned with the example presented by the Minnesota Department of Education in their Consolidated State Application.

**Figure 5. Straight line trajectory with plateaus.**

![Figure 5](image_url)

Note. This figure is based on the description presented by Porter, ct al. (2005) and Wiley ct al. (2005) which is aligned with the Illinois example for their initial and final years of AYP implementation in their Consolidated State Application.
Mathis (2004) calls it a fallacy to believe that schools have the ability to make steady progress on achievement tests. According to Mathis, the implementation of the line of trajectory by NCLB is not founded in any research. He argues that the implementation of the balloon effect (or back loaded trajectory as it is defined by Porter, et al. 2005) where there is a moderate amount of progress made initially and then a high level of progress in the later years, is only delaying the inevitable that all schools will eventually fail. Changes through the reauthorization of NCLB might be the only solution to the problem that eventually all schools will fail by 2014 (Mathis, 2004).

Cut Score Manipulation
Researchers in the field (Darling-Hammond, 2007a; Darling-Hammond, 2007b; Guilfoyle, 2006; Harris, 2007; Sunderman, Kim, & Orfield, 2005) call into question the use of a single measurement cut score with the analysis of school effectiveness. Harris (2007) examined the use of cut scores in the NCLB accountability measures using data sets from Michigan and Florida to illustrate how the different levels of cut scores can affect the number of students that are deemed high performing according to their poverty levels. Harris shows that when states use different cut scores like those in Michigan and Florida, the level that is used can directly impact the number of students that will be deemed proficient. Harris found that if a state uses a cut score that is halfway between the mean achievement of high minority students and low minority student populations, the result can be a dramatic shift in the numbers of students who do not make the proficiency level in the low minority schools. Harris discovered that states can manipulate the numbers of schools and students that are proficient by
simply changing the cut scores for the achievement of standards on their state assessments. The result is a dramatic impact on the proficiency averages of schools that have high percentages of poverty and minority students. The cut score manipulation that states might implement can provide the appearance that gains have occurred without any improvement in the conditions of school programs or classroom instruction. The findings by Harris identify another method in which states can use to modify the percentage of schools that meet AYP standards that is not a result of improvements in school quality. The manipulation of the statistical data can have a dramatic impact on AYP scores.

Lack of Consistency with AYP Across the Country

The implementation of the AYP accountability provisions in NCLB provide for a low level of consistency among the 50 states due to the wide variety of statistical methods that states use to calculate AYP. Porter, et al. (2005) analyzed how the state of Kentucky was implementing their AYP data according to subgroup size, confidence intervals, and the line of trajectory. They found that the state manipulation of those statistical measures in combination can have a dramatic impact on the number of schools that made AYP in Kentucky. For example, Kentucky holds schools accountable for subgroup achievement when they have a minimal of 60 in the subgroup, they use a two-tailed 99 percent confidence interval, and they have a back loaded trajectory. With the implementation of those measurements procedures, in 2003 the state of Kentucky had 90 percent of their schools meet the AYP requirements. In 2004, 94 percent of the schools met the AYP requirements under the same measurement procedures. When the researchers eliminated the use of the confidence interval, they determined that only 61 percent of the schools in Kentucky would have made AYP in 2003 and 72 percent would have made it in 2004. When they further created a straight line trajectory and eliminated the confidence interval, the researchers determined that only 45 percent of schools would have met AYP requirements in 2003 and 59 percent would have met AYP requirements in 2004. When they modified the minimum number of students for subgroups to be disaggregated from 60 to 30, used the straight line of trajectory, and eliminated the confidence interval, they determined that 31 percent of schools would have met AYP in 2003 and 44 percent in 2004. Figure 7 illustrates the data findings by Porter et al. with significant changes that occur as they changed the statistical methodology for the achievement of AYP in the state of Kentucky for the 2003 and 2004 school years.
Figure 7. AYP Proficient schools in Kentucky with statistical manipulation

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>90</td>
<td>61</td>
<td>45</td>
<td>31</td>
</tr>
<tr>
<td>2004</td>
<td>94</td>
<td>72</td>
<td>59</td>
<td>44</td>
</tr>
</tbody>
</table>

* Includes the use of the confidence interval, minimal subgroup size of 60, and a back end loaded trajectory.

* Includes the use of the minimal subgroup of 60 and the back end loaded trajectory with no confidence interval.

* Includes the use of the minimal subgroup size of 60, with a straight line trajectory and no confidence interval.

* Includes a minimal subgroup size of 30, with a straight line trajectory and no confidence interval.

Note: This figure is based on data that was presented in the study by Porter et al. (2005).

A number of implications have surfaced as a result of the inconsistencies in policy (Porter et al., 2005). The first implication is that there are different standards that are being implemented by different states in order to measure the AYP accountability provisions in NCLB. The result is little consistency between the states when it comes to setting AYP requirements. Another implication is that the different choices states are making with regards to how they design and measure AYP has a dramatic impact on the number of schools that meet AYP requirements. This includes making changes in the confidence interval measurement, the number of students in a specific subgroup, and the type of trajectory that a state employs. This means that it is virtually impossible to perform a fair comparison of the AYP data from state to state without examining the specific decisions that a state has taken while evaluating their schools according to AYP. It is clear that the statistical designs that states use to measure AYP have an enormous impact on the number of schools in a state that meet AYP requirements. Since failure to meet AYP comes with a cost through the implementation of sanctions that are required by NCLB for Title I schools, it is probable that states would continue to choose less stringent designs when measuring schools according to AYP.
Reliability Issues with a Single Measure Accountability System

There is a consensus among researchers in regards to the reliability issues that are associated with the use of a single measurement to evaluate school effectiveness. Several studies (Darling-Hammond, 2007a; Darling-Hammond, 2007b; Elmore, 2002; Harris, 2007; Hess & Petrilli, 2006; Kane, Douglas, & Geppert, 2002; Linn & Haug, 2002; Sunderman, et al. 2005) call into question the reliability of the single test method in evaluating school effectiveness. Elmore (2002) argues that there are many technical difficulties with the reliance on a single test to evaluate school effectiveness and the achievement of state curriculum standards. Elmore (2002) further claims that the creation of an accountability system at the state level was essentially a political act. He believes that the push for state accountability at the local level is a risky business for state politicians. Since most politicians do not have expertise when it comes to high stakes assessments, they are unaware of the reliability issues that are involved in using a single measurement to assess an entire school or district. Furthermore, Kane et al. (2002) point out a weakness with using a single year test to measure school effectiveness because different cohorts of students are measured against each other in order to determine a schools' AYP status. The study by Kane et al. emphasizes there are different variables at play within the cohort groups which ultimately have an impact on standardized test achievement. The result is unreliable data when it comes to measuring the effectiveness of schools.

High Flying Schools and the One Year Phenomena

Past research (Harris, 2007; Kane et al. 2002) found there can be fluctuations in annual test scores by schools in a single year that have nothing to do with gains in school performance. The study by Kane et al. (2002) found that the size of the school has statistical significance with the variance of test scores from year to year. The larger schools are more stable and have less variance over time. Thus, measuring students accordingly by only using year to year gains based on cut scores is statistically unreliable. “With test scores being so volatile, school personnel are at a substantial risk of being punished or rewarded for results that are beyond their control” (Kane et al., 2002, p. 60). Linn and Haug (2002) found that many schools that are being recognized as outstanding and others that are labeled as needing improvement which has to do with random fluctuations and not from changes in the quality of education in school performance.

Harris (2007) provides a direct rebuttal to research reports by the Education Trust and Heritage Foundation that have attempted to identify high-flying schools. High-flying schools are defined as those schools that help students reach high levels of achievement despite high poverty levels or high levels of minority students. Harris evaluated the data that was used by the Education Trust study to define schools that are labeled as high-flying. He analyzed the high flying definition that was used by the Education Trust at 1-1-1. This means that the school needs to achieve at a high level in either year of the study, in
any of two subjects, and in any of two grades that were measured. He then analyzed this data in a graphic form by making more stringent requirements with the use of the classifications 2-1-1, 2-2-1, and the 2-2-2. For example, the last definition 2-2-2 means that schools needed to perform at a high level for two years in two subjects and in two grades.

Harris (2007) found the data that was provided in the Education Trust study, which defines schools as high-flying, only took into account a single year’s worth of data. When multiple year data sets are used that take into account multiple tests at different grade levels, Harris found that low poverty schools are 22 times more likely to become high performing when compared with the high poverty schools. He also found that low poverty and low minority schools are 89 times more likely to become high performing vs. a high poverty and high minority school. The findings by Harris point out another problem with AYP as there is the possibility of the statistical chance where some schools will do well on a single assessment for an individual school year. However, the probability of achieving that same result for multiple years and over multiple tests is less likely to occur and the chance for statistical error is reduced.

**AYP Measures Social Capital, Minority Status, and Racial Subgroups**

Literature in the field (Elmore, 2002; Harris, 2007; Wiley et al. 2006; Meier, et al., 2004; Mathis, 2004a; Reeves, 2006) has shown that poverty and minority status have a major impact on the achievement levels of students in schools. Elmore (2002) points out that the implementation of AYP is doing nothing more than measuring the social capital of students in schools. Elmore argues that standardized test scores show the social capital of the students in schools. Most high-performing schools elicit higher performance by relying on the social capital of their students and families rather than on the internal capacity of the schools themselves. Most low-performing schools cannot rely on the social capital of students and families and instead must rely on their organization capacity. Hence, with little or no investment in capacity, low-performing schools get worse relative to high-performing schools.

The argument that Elmore (2002) frames leads to the conclusion that while the intent of NCLB is to improve schools for underperforming students; it will actually work in reverse as it broadens the gap between the schools with a high social capacity and a low social capacity. The justification for this argument is the primary focus of the NCLB reform which is the belief that school accountability measures and sanctions will motivate students and schools. Elmore indicates that this philosophy is based upon a faulty premise. The accountability sanctions do not provide organizational capital to the schools that require this capital in order to achieve a level of success with underperforming students.

The identification of what Elmore (2002) refers to as social capital is an interesting perspective because he argues in order to improve schools, there is a need to augment the social and organization cap-
maintains that the social capital of the students and their families is what is being measured by the standardized, single measurement accountability system. The result is a need for extensive studies in order to analyze his hypothesis along with an analysis of the possible positive and/or negative impact that the NCLB accountability provisions are having on public education in the United States.

Researchers (Mathis, 2004b; Meir, et al. 2004; Wiley et al. 2005) point out that schools cannot do everything to fix the societal problems. Mathis (2004a) argues that the schools cannot create other conditions in society where the students attend school with high levels of readiness, motivation, a high level of academic ability, and support from the home. He states that the use of test scores to close the achievement gap is more of an ideological belief rather than a belief that is founded in research. Mathis (2004a) further states that not all students start at the same level. He argues that it is unrealistic to believe that all students will be able to achieve the same proficiency levels. Mathis refers to personal capital which is the readiness and ability to achieve on tests. In conjunction with Elmore (2002), Mathis (2004a) argues that factors outside the school known as social capital have an impact on student achievement. He defines social capital as a combination of the following; parents’ education level, the values the family places on education, socioeconomic status, the peer group influence, and similar assets and liabilities. Mathis (2004a) goes on to say there is a great gap in the educational equity in schools in the United States and he cites a study by the Organization for the Economic Cooperation and Development which ranked the United States twenty-second out of twenty-five industrialized countries when it comes to educational equality. As schools are labeled as failing, it is possible that the NCLB reform will help to distinguish the disparities among the high poverty students and low poverty students.

Harris (2007) found that African-American kindergarteners are achieving at a rate of 34 percentile points below the levels of white kindergarteners. He believes that the current political climate has evolved in a manner in which school systems are blamed poor levels of student achievement. He states that the politicians are neglecting to look at other factors that come into play with regards to student achievement. Harris argues that a fault in NCLB is the failure to account for learning growth by students which is in contrast to the sole measure of cut score proficiency achievement. He asserts that the AYP measurement system cannot be solely attributed to learning in schools.

Harris (2007) illustrates his position with the following quote:

Consider the foot-race analogy made by President Lyndon Johnson when he argued for affirmative action and compensatory education. Johnson said that undernourished students would lose the vast majority of the running races, not because other students or track coach failed to try hard enough, but because the students were undernourished. (p. 373)
This analogy illustrates that there are other variables at play when students come to school with inadequate social capital. The problem with AYP is that it does not take into account the amount of social capital or the starting point at which the students enter school.

Inevitability of Labeling Schools as Failing

Wiley et al. (2005) found that a large number of schools will be labeled as failing by the year 2014. Their research is based upon an examination of six states in the Great Lakes Regions; 1) Illinois, 2) Indiana, 3) Michigan, 4) Minnesota, 5) Ohio, and 6) Wisconsin. The authors found that due to the different requirements and parameters that are mandated by AYP in each of the six states, each state exhibits a different but increasing pattern of having more schools not make AYP by the year 2014. This study did not take into consideration the implications of the subgroup issues or the integration of the science test into the AYP equation.

Even with the most conservative estimate, Wiley, et al. (2005) found that the number of failing schools is projected to total well over 50 percent. For example, in Michigan they found that in 2004, 25 percent of the schools failed to meet AYP. Michigan uses the stair step trajectory method until 2010 when it becomes a back-loaded trajectory. The minimum size for a subgroup is 30 and there was no confidence interval that was used by the state in 2005. According to the more conservative growth estimates, Michigan is predicted to have 3.2 percent of schools make AYP in 2014, while 46.2 percent will be safe harbor eligible, and 50.6 percent will not make AYP and will not be safe harbor eligible. With a more aggressive or low growth prediction, the averages in the same order change to 0.2 percent, 13.2 percent, and 86.6 percent. The sharp decline in schools that will not make AYP is predicted to occur in 2010 in Michigan due to the line of trajectory. Figure 8 illustrates the percentage of schools that are projected to make AYP and fail to make AYP in the year 2014 in Michigan.

The findings by Wiley et al. (2005) have major implications for the current state of
the NCLB accountability system. As the proficiency bar rises, a large majority of schools are predicted to be labeled as failing by 2014 under the NCLB reform. When looking at the AYP trajectory that leads to 2014, a reasonable assumption could be that the number of schools that fail to make AYP will increase similar to the predictions in this study. Looking at the number of schools that have currently failed to make AYP and aligning them with the trajectory increase by each state until the year 2014 may lead to this conclusion. The findings in this study show the only thing that may change that projection is revisions in Congress through the reauthorization process. Researchers in the field (CEP, 2005; Darling-Hammond, 2007b; Hess & Petrelli, 2006; Lewis, 2006; Linn, 2003; Mathis, 2004a) state that the target of having 100 percent of all students achieve high standards is statistically impossible and schools are set up for ultimate failure under NCLB. Mathis (2004a) references the fact that in 2003, 87 percent of schools in Florida and 57 percent in Delaware were labeled as failing schools. He concludes that an increasing number of schools will be punished as failing.

Summary of Measurement Concerns with the Use of AYP to Evaluate School Effectiveness

An analysis of the empirical research and literature in the field has provided a consensus that there are multiple problems with the NCLB accountability provisions. Every state has the autonomy to set up their own AYP proficiency standards and state administered tests in reading, mathematics, and science. The result is difficult to use AYP to accurately compare the quality of schools across the United States. The statistical manipulations of the AYP measurement systems at the state level provide the public with a false impression of a consistent measure of school effectiveness. The research leads to a conclusion that modifications to the accountability provisions in NCLB will be necessary in order for the legislation to effectively improve the conditions of school programs and classroom instruction.

AYP Benefits and Unintended Consequences

Philosophical intent

There is a consensus in the field (Borowski & Sneed, 2006; Darling-Hammond, 2007b; Guilfoyle, 2006; Haycock, 2006; Hess, 2006; Hess & Petrelli, 2005; Kane, et al. 2002; Lewis, 2006), that the most beneficial part of the NCLB reform is the philosophical premise that all children will count and no child will be left behind. Borowski and Sneed state the following:

In our view, NCLB's most important potential benefits lie in its recognition of the fundamental premise that all children can learn and have a right to be taught, and the corollary principal that educators, parents and students should receive periodic assessments of how students are progressing toward the attainment of high academic standards. (p. 504)

Borowski and Sneed's (2006) opinions have been molded through the litigation
of desegregation plans for schools in the United States. They often found that districts could not provide disaggregated data by race, students with disabilities, or the number of students in honors programs on state assessments. They assert that accountability systems have the potential to have a positive impact on public education as parents are informed about how schools are achieving on assessments and within particular subgroups. Borowski and Sneed (2006) state the following:

NCLB’s principal strength is its focus on improving students’ academic performance, and doings so for all children of all backgrounds. The act requires that all states establish standards in reading, mathematics, and science and develop assessments to determine whether students are proficient in meeting the standards. NCLB requires that all students meet these standards by the 2013-2014 school year. (p.506)

The philosophy found in NCLB is a focus on high academic standards and success for all students in all areas and subgroups which according to Borowski and Sneed (2006) have a great potential to provide educational equity. This potential is powerful because it mandates improvement not just in the aggregate, but also for all subgroups through disaggregated data. Borowski and Sneed argue that the intent of the legislation must go beyond having the data disaggregated in order to inform classroom instructional practices with the goal to make improvements for all students in schools.

Consequences of Sanctions and Accountability

Impact on Subgroup Populations

Haycock (2006) found that at Centennial Place Elementary School in Atlanta, administrators, and teachers are using data and information to assess the needs of students. The school disaggregated the data due to NCLB and they were able to pinpoint a weakness with special education students. She posits that the result was growth in mathematics and reading that produced a 30 point gain. She also references Woodville Elementary in Richmond, Virginia where there has been a 20 point increase in students meeting the standards on the state assessment. In order to support her conclusions from the schools that she references, Haycock refers to the Hall & Kennedy (2006) study which found that in elementary reading, 27 of 31 states showed improvement and 22 of 29 narrowed the gap between African American and White students while 24 of 29 narrowed gaps between Latino and White students. In elementary math, 29 of 32 states showed improvement, 26 of 30 narrowed the gap between African American students and White students, while 22 of 30 narrowed the gap between Latino and white students.

Haycock (2006) claims that a 30 point growth was made by a school in Atlanta as a result of the positive impact that NCLB is having on the school. However, there are numerous variables that could have had an impact on the achievement of that school. Some of the previous literature has illustrated the manner in which the
states manipulate their variables like subgroup (N) totals, confidence intervals, and the proficient cut scores on assessments, which could also have an impact on the annual gains that are referenced.

**Poverty and racial diversity**

Research in the field (Mathis, 2004a; Meir, et al. 2004; Sunderman, et al. 2005) shows that the poor and ethnically diverse schools will be punished first through the sanction provisions. Those schools with more subgroups and higher percentages of racial groups of students will have a greater chance of failing to meet AYP standards which will result in the diversion of funding from those schools to pay for the mandated sanctions. Porter et al. (2005) posit schools with a 25 percent or more black student population in Kentucky were much more likely to fail AYP at a 25-30 point increase. They also discovered schools with higher percentages of Economically Disadvantaged (ED) populations were more likely to fail to meet the AYP requirements. When they developed a demanding AYP design by eliminating the confidence interval and using a straight-line trajectory, there were an even greater number of schools (73 percent) with a high ED population that failed to make AYP. This was in contrast to only 30 percent of schools failing to meet AYP requirements with lower ED populations. A study by Springer (2008) found that schools with larger Hispanic populations were 8 percent more likely to fail AYP. Schools with larger ED populations were more likely to fail to meet AYP by 9 percent. However, it is difficult to reliably conclude that these results are transferable to other states since there were lower numbers of racial subgroups and lower numbers of ED students in the state that Springer studied when compared to other states across the country.

**Educational Triage**

Educational triage involves focusing on the students who are just below the proficiency level in order to accelerate their achievement so that they meet the state standards. Those students are often labeled as the “bubble kids”. Guilfoyle (2006) argues that due to the NCLB accountability provisions, schools are focusing on the students who are achieving just below the cut line in order to improve their high stakes test scores. Springer (2008) suggests that if educational triage is occurring in schools, then the students in the low categories that have little chance of passing along with those in the high categories that have already achieved proficiency might receive less attention and resources. The “bubble kids” benefit by receiving resources that might have otherwise been used for the highest or lowest achieving students.

The findings by Springer (2008) indicate that educational triage is not occurring in schools. Springer analyzed the test scores from over 300,000 students from a western state in the United States. He used the academic achievement data from the fall and spring state test scores in the area of mathematics. The state that he used for his analysis has two assessments that are given to their students in one academic year as opposed to the one assessment that is mandated by NCLB. Springer could not
find any evidence that schools were involved in educational triage with a reallocation of their resources to the bubble kids. Some of his findings conclude that the lowest students gained more than those near the proficiency line for the schools that failed to make AYP during the previous school year. There was a 0.2 standard deviation gain vs. the students who were near the proficiency line. The students in the highest categories did not have a negative consequence as their achievement was also in line with the “bubble kids”. With the schools that made AYP the previous year, Springer found that the students in the lowest category the previous year showed the highest gains. He also found that proficient students in failing schools gained more than proficient students in non-failing schools. These results suggest that the higher achieving students and lower achieving students are not losing out on educational resources or interventions for the benefit of the “bubble kids” so that the schools can make AYP. However, since there are very small numbers or racial subgroups in the state that was studied, it would be difficult to conclude that these results were prevalent in other states with larger populations of racial or ethnic diversity. One weakness with the study is that Springer only used mathematics test data and not language arts or reading test scores which is an equally important part of the AYP accountability formula. As a result, the findings can only be attributed to mathematics instruction in the sample state.

Narrowing the Curriculum

Scholars posit that NCLB is narrowing the curriculum as it discourages the usage of instructionally useful forms of assessment that involve extensive writing and analysis (Darling-Hammond, 2007a; Ravitch, 2010). Some states have been forced to eliminate the usage of sophisticated performance-based assessment systems that resemble those used in other nations around the world that score well on international assessments. Guilfoyle (2006) argues that testing and measurement is a major emphasis in NCLB and since mathematics and reading are being tested, there is an unprecedented focus on those two areas. The areas of the curriculum that are not being measured by NCLB are suffering as schools start to preclude or reduce their focus in those areas. This includes history, art, civics, music, physical education, health, and other cultural areas (Ravitch, 2010). Guilfoyle (2006) argues that although these areas are not being measured, they do have a tremendous impact on the education of students and they are extremely important. Since those areas are not being tested, students and teachers are not placing much emphasis on those content areas. Guilfoyle (2006) also indicates that standardized assessments provide only a limited amount of information. The testing provisions in NCLB are unable to provide information about prevalent instructional improvement. But, empirical data was not presented in order to determine if there actually is a narrowing of the curriculum. However, since NCLB measures growth in mathematics and reading, it is probable that schools would place more
emphasis on those subjects while reducing the other areas of the curricula. Sunder-
man et al. (2005) collected qualitative survey data from teachers in two school
districts Fresno, California and Richmond County, Virginia. They found that teach-
ers acknowledged that there was a narrowing of the curriculum as they indi-
cated that more focus was given to the test areas in contrast to those areas that were
not tested and could be labeled as a form of curricular triage.

Focus on Basic Skills vs. Higher Level Thinking and Innovation

A strategy that educators could use to improve tests scores on the NCLB account-
ability measurements is to teach to the test where the sole focus is on basic skills and
test taking methods. However, teaching to the test will not improve the schools over-
all educational program in order to prepare students for society. Schoen and Fusarel-
li (2008) along with Guilfoyle (2006) argue that there is a need to create a system that
rewards educators for innovations in schools instead of the current system that places the utmost importance on mastery of core content and basic skills. Schoen and Fusarelli (2008) believe that the constructivist approach that supports the acquisition of the skills that are needed for students to become successful in the 21st century are in competition with the accountability provisions in NCLB that emphasize content based learning and the need to do well on a single measure. Schoen and Fusarelli (2008) along with Darling-Hammond (2007a) claim that there is a need to change the current assess-
ment system that is supported by NCLB. They believe that new assessments must be
designed through the reauthorization of NCLB. Those assessments should mea-
sure higher order critical thinking skills, problem solving, and the ability to process information in a relevant manner. The NCLB legislation has narrowed the cur-
riculum and essentially made it difficult for schools to promote innovation and move towards the ideals of the 21st century schools movement (Ravitch, 2010; Dar-
ing-Hammond, 2007a; Darling-Ham-
mond, 2007b; Meir et al., 2004; Schoen & Fusarelli, 2008).

Recommendations for AYP reform

The Need for Empirical Evidence that Supports the Use of the AYP Accountability System

The intent of the NCLB reform is to improve classroom instruction and the educational programs for disadvantaged students. The consensus found in the liter-
ature suggests that policies aligned with the AYP provisions in NCLB are not com-
pletely founded in empirical research given the goals of the reform. They were creat-
ed on a political or ideological basis instead of a scientific research basis. For exam-
ple, many of the findings and conclusions that have been presented in this review (Borowski & Sneed, 2006; Elmore, 2002; Harris, 2007; Kane, et al. 2002; Lewis, 2006; Sunderman, et al. 2005) indicate that using a single once a year high stakes test for an accountability measure is unreliable in evaluating school effectiveness. Sun-
derman et al. (2005) found there is a lack
of empirical data, which shows that the school of choice provisions or supplemental service provisions will produce improvements for students. They point to the fact that the sanctions might actually weaken the conditions of students in schools as funds are diverted.

Elmore (2002) states that if there is a lack of support given to the internal capacity of schools to make improvements that the sanction provisions will have little impact on improving classroom instruction as schools are labeled as failing according to AYP. Schoen and Fusarelli (2008) interviewed educators and they found that the accountability sanctions also created pressure on principals who considered the possibility of leaving schools due to sanction provisions. Marzano, Waters, and McNulty (2005) found that school leadership has a substantial impact on student achievement. If high quality principals leave schools due to the pressures of probation, then it is probable that those schools will have less of a chance of making improvements.

Borowski and Sneed (2006) believe that the accountability provisions and sanctions in the NCLB legislation are unproven and lack scientifically backed data that supports the implementation of those provisions. In their estimation, the legislation is a bandage job that attempts to balance between implementing federal intervention and allowing for local and state control of education. Thus, they observe the potential for grave harm as states have at times lowered standards or manipulated their statistical data in order to avoid the costly penalties and sanctions.

Lewis (2006) argues there are major problems with the current model of accountability known as AYP and believes it needs to be eliminated and a new model should be grounded in accountability research that will help to improve the conditions of schools.

The Use of Growth Data for AYP

A large number of empirical research studies and articles (Elmore, 2002; Hess & Petrilli, 2006; Koretz, 2008; Peterson & West, 2006; Popham, 2005b; Schoen & Fusarelli, 2008; Sunderman et al. 2005; Wiley et al. 2005) recommend the use of growth data in the AYP formula. Research by Sunderman et al. (2005) found that the state accountability systems that were developed in each of the states were generally in contrast to the accountability provisions in NCLB. Most of the states used a form of growth data in order to evaluate the schools in their states. They did not solely rely on the absolute cut score method that is prevalent in NCLB. This often creates confusion for school stakeholders. For example, the schools in Chicago were subject to three accountability provisions at the national, state, and local levels — all of which could create different results from the measures that were implemented in each system.

Peterson & West (2006) used the Florida Comprehensive Achievement Tests (FACT) for their study for the 2002-2003 and 2003-2004 school years. They found that schools making AYP had a standard deviation that was 9 percent above the amount gained by students at schools not making AYP. This standard deviation sta-
tistic was stated to be equivalent to 1/3 of a school year of student growth. The achievement in reading was a 7 percent standard deviation higher in the schools that made AYP when compared to those schools that did not make AYP. In schools that made AYP, the standard deviation was 11 percent higher for African Americans and 12 percent higher for Hispanic students when compared to the same subgroups of students in schools that did not make AYP. There was a 6 percent difference of a standard deviation in the area of reading for both of those ethnic groups. When they completed a statistical analysis to take into account for the ELL, special education, socioeconomic status, ethnicity, and mobility rates, the average was 4 percent of a standard deviation in mathematics and 2 percent of a standard deviation in reading for schools that met AYP vs. those schools that did not meet AYP.

Schools in Florida that made AYP outperformed the schools that did not make AYP 71 percent of the time according to the analysis of student growth data on the mathematics assessments. This resulted in a 28 percent error rate. Peterson and West (2006) disaggregated the data to further analyze the differences according to growth data in relation to the Florida system that evaluates the schools on a 5 point scale A through F. While the learning gains were at a 7 percent standard deviation higher in schools that scored A versus those that scored F, the gains between the A schools and the F schools was 25 percent of a standard deviation or the equivalent of one school year of growth. In 2004, only 47 of the states 2649 schools were given an F and 184 were given a grade of and E. This is in contrast to a total of 75 percent of the schools that did not make AYP including more than half of the schools that received an A grade.

The findings by Peterson and West (2006) have important implications as the reauthorization of NCLB is being debated in Congress. The study shows that the AYP data does not have the ability to measure the growth that is occurring in schools. It brings into question the true goals in NCLB. For example, would a parent want to send their child to a school that is having more success with student gains vs. those that appear to be making the cut scores in NCLB? The study points out that the accountability system in Florida and the NCLB accountability system creates confusion for Florida residents. Half of the schools that scored an A on the Florida accountability system are labeled as failing under the NCLB accountability system. The data further suggests that AYP has a large error rate when labeling schools as successful in accordance with the growth data.

A Call for Multiple Measures

The use of a single measure with the evaluation of school effectiveness is not a reliable source of measurement. Several researchers (Darling-Hammond, 2007a; Elmore, 2002; Guiffoyle, 2006; Kane et al. 2002; Scheon & Fusarelli, 2008; Wiley et al. 2005) support the use of multiple measures in order to evaluate school effectiveness. Kane et al. (2002) examined the impact of the use of a single measure to evaluate school effectiveness in their study.
They used data from California to show that schools with racially homogeneous populations have a better statistical chance of winning performance awards under the subgroup accountability provisions in that state. They point out that schools with larger racial subgroup populations have a greater chance of failure and that the NCLB reform might provide an incentive for states to lower standards so that more students meet the proficiency levels. The authors identify four different types of accountability systems that were used across the country prior to NCLB. States like Arizona, North Carolina, and Tennessee used a growth model. States such as Texas and Illinois reverted to a cross cohort comparison where different cohorts of students were compared from one year to the next. Other states like California rated their schools based on change in test scores from one year to another. Some states used accountability methods that looked at including a combination of growth and proficiency cut scores from year to year.

Researchers (Darling-Hammond, 2007a; Elmore, 2002; Guilfoyle, 2006; Scheon & Fusarelli, 2008) further argue that authentic learning projects and portfolios should be emphasized at a higher or equal level in NCLB. An emphasis on those types of assessments with multiple measures would promote higher level thinking and critical problem solving approaches to education. Another form of measurement that could be used for the evaluation of school effectiveness is a process that has been implemented by the Advanced Education Organization, formerly known as North Central Accreditation (NCA). This process involves the evaluation of school improvement plans, the inclusion of growth data, the use of multiple forms of achievement data, and the reliance on school visits to measure school effectiveness to determine accreditation status. While this measurement would be costly, it would be beneficial to explore this option as a part of the AYP measure with regards to school effectiveness.

National Curriculum and the Use of the NAEP

AYP provisions are not consistent from state to state as different states have different cut scores and their trajectories are often much different (Sunderman et al., 2005). Research shows that even when one state might have higher standards when compared to another state based on the proficiency levels, the proficiency targets could be deceiving because the quality of the assessments used might be much different in each state. Borowski and Sneed (2006) argue that there is a need for national standards that protect the local rights and state rights for implementing educational policies and that minimal national standards are necessary. They emphasize that the arbitrary accountability provisions throughout the AYP process must change in order to equalize the accountability standards among the states in a fair manner. Hoxby (2005) believes there is need to create a national benchmark system in order to effectively evaluate school progress across the country. Since the legal authority to implement education falls to state
governments, it is predicted to be difficult in getting approval in Congress and consensus among the states for a national curriculum standard that is used in all states. It is also probable that there would be court challenges to any form of national curriculum that is mandated by the federal government. However, as Ravitch (2010) points out, we should establish a national curriculum that "declares our intention to educate all children in the full range of liberal arts and sciences, as well as physical education" (p, 231).

What We Still Need to Know

The current accountability provisions in NCLB have not proven effective in evaluating school performance. It is apparent from the literature that there are numerous problems with the implementation of AYP in order to motivate effective school improvement efforts. Reforms to the legislation are needed in order to improve the reliability of measuring school effectiveness. This includes the need to standardize the AYP measure across the country while implementing forms of measurements, which include growth data. Since the key to school improvement is through the enhancement of the internal capacity of the school while increasing the internal accountability mechanisms, it is critical that research is conducted in this area. Research is needed regarding the impact that AYP is having on student learning, classroom instruction, and school improvement. Future work should examine the impact AYP is having through other measures, including qualitative data analysis that examines how teachers and administrators are responding to AYP. Moreover, research is needed to address the question of whether a school will improve if the staff is replaced? Will a school get better if it is turned into a charter school? Will it improve in measureable ways if the state takes it over? What policies could be implemented to attract and retain high-quality teachers in low performing urban schools? These and a good many other questions remain to be answered within the context of NCLB.

References


