Running head: Equity in Accountability Systems

Promoting Equity in State Education Accountability Systems

Chris Domaleski

National Center for the Improvement of Educational Assessment

Marianne Perie

Center for Educational Testing and Evaluation, University of Kansas

Abstract

This paper describes state initiatives to improve school accountability systems in response to the federal Elementary and Secondary Education Act flexibility initiative. Of particular interest is the degree to which system designs may promote improved academic achievement for low-performing students, broadly termed "equity." Accountability systems that support equity are those that are more effective at incentivizing actions that lead to academic improvement for the lowest performing students and detecting these desired outcomes. Four prominent state accountability initiatives are explored in detail: consolidated subgroups, achievement gaps, growth, and mechanisms for combining measures. These accountability measures present opportunities to better include all students in the accountability system, but they also pose threats to the promotion of other equity considerations. Therefore, suggestions are presented to help guide evaluation and monitoring of these practices to deepen the understanding of accountability design features that support equity.

Introduction

The landmark Elementary and Secondary Education Act (ESEA), which was signed into law in 1965 as part of President Johnson's war on poverty, represented a sweeping initiative to improve the equality of educational opportunities for public school students in America. This legislation, "focused attention on the educational needs of poor children and established federal standards to push school districts toward more equitable treatment of disadvantaged students," (Kantor, 1991). In the nearly 50 years following the ESEA, the pursuit of equity, perhaps more than any other goal, has dominated educational policy and reform initiatives.

The term equity is used throughout this paper to broadly refer to the aim of promoting improved academic achievement for low-performing students, particularly those deemed below a defined standard of achievement (e.g., below grade level or not proficient). Equity is based on the principles of fairness and justice, and holds that unequal access to learning opportunities leads to gaps in academic achievement. Numerous works have documented persistent gaps in educational opportunities and achievement-based factors such as race, economic class, and geography (Lee, 2004; Darling-Hammond, 2007).

The earliest equity initiatives following the passage of ESEA typically focused on inputs (Fuhrman & Elmore, 2004). That is, schools were largely held accountable for providing adequate resources and complying with regulations. This focus began to shift during the 1980s in the wake of concern about the perceived decline in quality of education described in the influential publication *A nation at risk* (National Commission on Excellence in Education, 1983). Subsequently, standardized achievement testing became more commonplace and was extended in the 1990s with increased support for standards-based reform (Goertz, 2005; Lee & Wong,

2004). The 1994 reauthorization of ESEA, called the Improving America's Schools Act (IASA), increased Title I funding for schools serving low-income students, with the stipulation that these schools comply with more federal regulations, including standardized testing for students served by the program. The 1997 Individuals with Disabilities Education Act (IDEA) further codified the federal policy emphasis on access to educational opportunities and the importance of evaluating student performance against well-defined academic standards.

The 2001 reauthorization of the ESEA, No Child Left Behind (NCLB), introduced a much stronger emphasis on outcomes-based school accountability to support the legislation's now titular goal of equity. NCLB broadened the equity expectations from low-income students to other low-achieving student groups. This marked the culmination of a gradual change from earlier accountability initiatives and an explicit endorsement of universal proficiency. The NCLB approach to equity was clear, if somewhat controversial. NCLB required three main components:

- States must adopt academic standards in reading/language arts, mathematics, and science and assess these standards annually.
- Assessment results must be publicly reported and disaggregated by each identified subgroup to include: economically disadvantaged (ED), English language learners (ELL), students with disabilities (SWD), identified ethnic groups, gender, and migrant status.
- States must make annual determinations of adequate yearly progress (AYP), with the goal of having all students in each subgroup proficient in reading and mathematics by the 2013–2014 academic year.

These elements, among others, provided the basis for a very strong system to explicitly incentivize equity outcomes by holding schools accountable for the proficiency of all students.

Perhaps not surprisingly, criticisms of the NCLB accountability approach increased over time. One contentious issue was the status-based proficiency criterion. Critics argued that holding all students to a single threshold for performance was too coarse and failed to consider the academic growth demonstrated by low-performing students (Ho, 2008). Another common criticism was that requiring each subgroup to meet performance targets in order for the entire school to make AYP (i.e., conjunctive combination of components) was not advisable (Goertz, 2005). The logical outcome was that more diverse schools were much more likely to fail to meet AYP because of the sheer number of performance thresholds for which the school was accountable, typically between 35 and 42. Some argued that this issue was augmented by the fact that NCLB subgroups as defined are not mutually exclusive; students are often classified in multiple groups (e.g., a student may be classified as Hispanic, ED, and ELL). Finally, and perhaps most prominently, many felt that NCLB's ultimate requirement that 100% of all students—at the school level and within each subgroup at the school—must be proficient by 2014 for the school to meet AYP was an unreasonable goal (Rose, 2007). This prompted many critics to argue for a new definition of AYP (Rose, 2007).

In light of these and other concerns, state leaders worked with the Council of Chief State School Officers (CCSSO) in 2011 to develop principles for a next-generation accountability system. The result was an affirmation that every student should have access to a high-quality education system and that accountability systems need to focus on providing timely, transparent data that distinguish performance in a way that allows policymakers to target appropriate supports. The next-generation systems that these state leaders envisioned would allow for greater flexibility and innovation in accountability components and design. For example, systems should

be permitted to incorporate a more authentic measure of academic growth and broader set of performance indicators.

Even as concerns about NCLB increased and researchers identified improvement options, there was no legislative action to reauthorize ESEA. Therefore, in September 2011, U.S. Secretary of Education Arne Duncan announced that the U.S. Department of Education (USED) would provide some flexibility for the NCLB mandates through a series of waivers. In order to be eligible for those waivers, states had to meet three core principles.

First, each state was required to demonstrate that it had college- and career-ready expectations for all students. Readiness, in lieu of proficiency, became the primary outcome of interest and states were required to show that standards and assessments would be aligned to this new target. Most commonly, states addressed this by adopting the Common Core State Standards (CCSS) and by joining one or more of the Race to the Top state consortium assessment programs; either the Partnership for Assessment of Readiness for College and Careers (PARCC) or the Smarter Balanced Assessment Consortium (SBAC).

Second, states were required to develop differentiated recognition, accountability, and support programs. The systems were required to address improving academic achievement, closing achievement gaps, and setting ambitious but attainable annual measurable objectives (AMOs). It was this principle that gave rise to new equity approaches that are the primary focus of this paper. Many states responded by developing systems that redefined the way subgroups were identified, addressed achievement gap closure, prominently featured "true growth," and combined indicators in a compensatory rather than conjunctive manner. These measures will be discussed in more detail in subsequent sections.

Third, states were required to support effective instruction and leadership, drawing on multiple measures to evaluate teacher and leader effectiveness. This involved making a commitment to develop an educator evaluation system that meaningfully differentiated performance and included data on student growth.

It should be noted that new approaches permitted under ESEA flexibility waivers for establishing AMO targets generated no small amount of controversy. Particularly contentious was the widespread approach of setting different academic targets for different subgroups, such as linking AMOs to a reduction in the percentage of students not proficient over time (typically to reduce the percentage of nonproficient students by half in six years). In fact, an analysis of state waivers by *Education Week* revealed that of 34 waivers, only eight states proposed to set the same target for all student groups (McNeil, 2012). Although this is not significantly different from the application of safe harbor[1] used by most state NCLB plans prior to the waiver, critics charge that explicitly differentiating goals by subgroup works against the principles of fairness and equity. Although an analysis of this practice is beyond the scope of this paper, it is mentioned to illustrate the tension that policymakers face in accountability design to determine ambitious and attainable academic goals for all students.

Much has changed in the nearly 50 years since the passage of ESEA, but the focus on equity remains. States are engaged in new efforts to track different kinds of data and to use results to direct resources and inform initiatives. It is critical to pause and examine these practices to better understand the challenges and opportunities that each presents, and to identify the kind of monitoring and evaluation that is likely to inform understanding about the promising practices of the future.

[1] Safe harbor refers to the common practice of counting a subgroup as having achieved the AMO target if the percentage of nonproficient students in that group is reduced by 10% from one year to the next.

Purpose and Method

Given the prominence of equity concerns in education policy, this paper aims to provide an overview of contemporary approaches that emphasize equity in state accountability systems. Specifically, this paper will focus on approaches prominently featured in state responses to the federal ESEA flexibility initiative: consolidated subgroups, achievement gaps, growth, and mechanisms for combining measures. The authors reviewed all submissions that were approved by the USDE as of August 2012 and categorized approaches used by each state by multiple issues, including subgroup inclusion, growth measures, achievement gap calculations, and final decision metrics. Summary information is presented to gauge the extent to which these approaches are being used and to broadly identify similarities and differences among state practices in the 34 state proposals studied for this paper. The remainder of the paper is divided into four sections focusing on each of the aforementioned accountability approaches . Suggestions to help guide evaluation and monitoring of accountability practices are discussed within each section, acknowledging that these accountability mechanisms present opportunities to include all students in the accountability system and pose threats to other equity considerations.

Current Status of State Accountability Systems

As of August 2012, 33 states and the District of Columbia (hereafter referred to as a state, bringing the total to 34 states[2]) had received approval of their flexibility plans and had many restrictions waived, most importantly the requirement to have 100% of students proficient by 2014. The authors reviewed all 34 approved submissions and analyzed the approaches to various

[2] In October 2012, Idaho was approved as the 35th state. This paper was being finalized at the time, so no information from Idaho's proposal is included.

components of the accountability section, focusing specifically on those related to equity issues. Areas of commonality were noted and categories emerged for each approach. In addition, unique approaches were identified. Summaries were created for each state across various categories. This paper reviews the use of consolidated subgroups, metrics used to calculate achievement gaps, and features of the growth model as measures to maintain equity among student groups.

Consolidated Subgroups

Although every state is required to maintain the ESEA subgroups for purposes of calculating AMOs and reporting results, they are allowed to combine the groups for purposes of making accountability decisions. More than three-quarters of the states who had an approved flexibility request used some form of a consolidated subgroup (Table 1).

Table 1

Number and Percentage of States Using Consolidated Subgroups by Type of Consolidation

Type of Consolidation	Number of States	Percentage of States
None	8	24%
Lowest quartile	7	21%
Combine at-risk ESEA subgroups	6	18%
Only combine groups if minimum *n* is not met[a]	3	9%
Bottom 30%	3	9%
Below proficiency for growth	2	6%
Other	5	15%

[a] One state only combines three groups if minimum *n* is not met; another consolidates nine subgroups into three; the third uses different strategies depending on the indicator.

The most common approach was to examine the lowest performing students separately, regardless of their student group. Seven states focused on the bottom 25% while three focused on

the bottom 30%.[3] The arguments for focusing on the lowest performing students include being able to target resources to all students who are most in need and being able to count and include more students because the requirement of reaching a minimum number is lifted. Additionally, states typically provided tables showing that they would capture more students in each subgroup, because for small schools, the number of students in each subgroup often did not rise to the level needed for reporting, usually 20 to 30 students per category in most states. For example, Indiana chose to focus on the lowest quartile (Table 2). With a minimum number of 30 students needed to include each subgroup in an accountability calculation, a much greater proportion of students are included in the calculations using the lowest quartile. For example, under the traditional ESEA requirements, only 57% of schools are held responsible for the achievement of special education students as a separate category, although they are included in the "all students" group (Table 2). However, when the focus shifts to the lowest 25% of performers, 99% of schools are accountable for the performance of at least one student with disabilities in the separate subgroup. Naturally these findings were often most pronounced in states with a large number of schools that have little variability in subgroup membership.

[3] One state focused on the bottom third, but it was counted along with the states who specified 30%.

Table 2

Percentage of Indiana Schools Accountable for Student Performance in Each Subgroup under Two Conditions

ESEA Subgroup	Using Traditional Approach with Minimum *n* of 30	Using Bottom 25%
American Indian	0%	16%
Black	23%	62%
Asian	3%	31%
Hispanic	22%	71%
White	91%	97%
Free/reduced price lunch	90%	99%
Limited English proficient	19%	59%
Special education	57%	99%

However, civil rights groups and special education advocacy groups were concerned that specific groups of students may continue to struggle but have their low performance obscured because they are part of a larger group. Critics argue that requiring states to report out individual subgroups without requiring them to use that information to identify struggling schools or allocate resources could result in schools evading accountability for the performance of small, low-performing groups.

The second most frequent consolidation approach was to combine all at-risk groups. Typically, this included Black, Hispanic, low-income, ELL, and SWD categories. States using this method argue that by measuring the performance of the aggregate of all at-risk students, they "are able to hold more schools accountable for necessary progress in these high needs areas" (from the Missouri proposal, downloaded 9/5/12 from https://www2.ed.gov/policy/eseaflex/approved-requests/mo.pdf). In some states, such as

Wisconsin, the consolidated subgroup is only used in schools where one or more subgroups fail to meet the minimum number of students in order to be counted separately. However, once states determine consolidation is needed, they use a different type based on purpose. For schools with subgroups that do not meet the new minimum *n*-size of 20, they create a "super subgroup" of all at-risk groups to make initial determinations on meeting the AMOs. However, these states use the lowest 25% in their index. For achievement gap calculations, they examine White vs. non-White; SWD vs. non-SWD; ELL vs. non-ELL, and ED vs. not ED.

In other states, such as Massachusetts, policymakers chose to both lower the minimum group *n*-size and also use a consolidated subgroup. This "high needs" subgroup is comprised of students who are low-income, have a disability, or are ELL or former ELL. In lowering the minimum *n*-size from 40 to 30, Massachusetts will hold more than 100 additional schools accountable for students who are English learners, have disabilities, or who come from low-income families. Then, by using the high needs subgroup for accountability purposes, an additional 200 schools that currently do not have sufficient numbers of students in those individual categories will now be held accountable for the performance and progress of those students.

Other approaches of consolidation were also used. For example, Oregon combined all non-White races together but kept all other ESEA categories separate. In Rhode Island, when the sample sizes preclude a group from being counted, schools will combine ELL with SWD groups; combine all non-White students together; and combine low-income students with non-White students as needed to ensure everyone is included in the accountability calculation. Virginia also focuses on three groups, combining ELL and SWD, Black and Hispanic students, and focusing on low-income students as the third group.

Minnesota weighted accountability determinations proportionally to the size of the sample. This feature could have significant effects for schools with high-needs populations that are large enough to be counted but smaller than the majority. Consider, for example, a school that has 170 White students and 30 Black students. There are enough Black students to be counted for accountability given a minimum *n* of 30. Under the AYP conjunctive model, the school would be equally responsible for both groups. Under the new flexibility model, however, White students make up 85% of the population, while Black students make up 15%. Using the square root to determine the weights, the scores of White students are multiplied by 9.2, while the scores of Black students are multiplied by 3.9, less than half the weight of White students.

The National Center for Learning Disabilities (NCLD) wrote a letter to Secretary Duncan sharing their concerns with any approach that combined student groups, claiming that "combining the performance of several student subgroups does nothing to help schools identify how to go about targeting instruction to the students who comprise the group," (2012) They noted that reducing the minimum *n*-size for inclusion would have similar effects and should therefore be considered in lieu of consolidating subgroups. Several states did, in fact, reduce their minimum *n*-size. For example, Delaware, Massachusetts, and Mississippi lowered their minimum *n*-size from 40 to 30; Arkansas from 40 to 25; Washington from 30 to 20; Connecticut from 40 to 20; and the District of Columbia from 25 to 10.

Achievement Gaps

The achievement gap measure is an option for determining which schools should be categorized as focus schools[4] requiring targeted interventions. Although at least three states

[4] A "focus school" is a Title I school in the state which, based on the most recent data available, is contributing to the achievement gap in the state. The total number of focus schools in a state must equal at least 10% of the Title I schools in the state.

chose to identify schools with the lowest performing subgroups and without calculating a gap score, the majority of approved state applications included an achievement gap measure.

The instructions in the proposal guidelines asked states to compare the lowest performing subgroup to the highest performing subgroup. Similar to NCLB, this approach leads to nonmutually exclusive groups. For example, if the lowest performing subgroup is SWD and the highest performing subgroup is White students, there will be White students with disabilities counted in both groups. Although three states followed those instructions exactly, the majority of states proposed solutions to ensure that comparison groups did not overlap so that each student is counted only once.

In addition, the majority of states addressed the gap in a way that ensured comparisons were linked to the groups regarded as high-priority. As described above, the USED proposal could lead to comparisons that are not a priority focus for states, such as SWD compared to Asian students. The Collaboration to Promote Self-Determination (CPSD), a disabilities advocacy group, also addressed this issue in their policy brief, stating:

> We are concerned that this formula may not identify the achievement gaps that should be closed. The issue is not how subgroups are achieving with respect to each other, but rather how they are achieving with respect to all the other students who are not in that subgroup. For example we are interested in the gap between students with disabilities and students without disabilities, not between students with disabilities and English language learners. (2011)

Kentucky and Tennessee adopted a position similar to CPSD's by keeping each group separate but comparing it to its counterpart to ensure no students were double-counted. That is, Whites are compared to non-Whites, low-income to moderate- or high-income, SWD to students

without identified disabilities, and ELLs to non-ELLs. Tennessee went a step further by weighting each gap by the percentage of students negatively affected by it. That is, the Black/White gap was weighted by the percentage of Black students in the school, and the disability/nondisability gap was weighted by the percentage of students in the school with disabilities. Other states with a consolidated subgroup typically compare that subgroup to its counterpart so that there are no duplicated counts on either side of the equation.

Two states compare the school to state gap, focusing on the difference in proficiency rates between the school's subgroups performance and the state's all-student performance. New Jersey averages the percentage of students proficient in the two lowest performing subgroups in each Title I school. Then, that percentage is subtracted from the percentage of proficient students in the highest performing subgroup. To be included in this analysis, a subgroup must have a minimum size of 30 students and represent at least 5% of the total student population.

For states that chose to focus on the bottom quartile as their consolidated subgroup, the reference group was the top quartile for some, the top half for at least one, and the top 75% for others. Michigan compares the bottom 30% to the top 30%.

The majority of states use percentage of students at proficient or above as the metric for achievement gap comparisons (Table 3). There are, however, a few exceptions. Colorado focused on the growth of students and compared the median growth percentile (MGP), a normative measure, to the adequate growth percentile (AGP), a criterion measure, for each subgroup. When the MGP exceeded the AGP students were considered to be on target to reach or remain at proficient or above. Indiana also focused on the growth gap of the bottom 25% of students to the top 75%. Some states calculated the gap in their index scores with the index weighting performance levels and growth scores differentially. In another approach, Georgia

converted their scale score to *z*-scores and compared the highest and lowest performing subgroups on the *z*-score scale. Four states simply subtracted the average scale scores of the two groups and then analyzed the size of the difference.

Table 3

Number and Percentage of States Using Each Metric to Calculate Achievement Gaps

Metric	Number of States	Percentage of States
Percent proficient	18	53%
Growth gap	5	15%
Scale score	4	12%
z score	1	3%
Index gap	1	3%
Other	1	3%
None	4	12%

The achievement gap is the measure for focus schools in some states, and in others it is part of an overall index or school grading system. How the gap is portrayed and applied, combined with other measures, is an important component of equity. This issue will be discussed further in the section on combining indicators.

Growth Models

Each state's proposal included a plan for measuring student growth over time. Although some states examined the growth for individual subgroups and others examined growth for consolidated subgroups, combining growth with the percentage of students reaching the proficiency target was a relatively new metric for making accountability decisions. Under the growth model pilot program, states had to include a model that only counted students as meeting the growth target if they showed they were on track to reaching proficiency within three years. The directions for the flexibility waiver proposal included no such requirement, and instead

directed states to show how they would combine status (percentage proficient) and growth to determine whether schools were meeting their performance targets. States had to clarify how targets were set. Table 4 shows the distribution of states using various types of growth in the accountability system. It is important to note that some states chose to use a different model of growth for teacher evaluation; this paper only addresses growth used to make school-level classifications.

Table 4

Number and Percentage of States Using Various Growth Models

Growth Model	Number of States	Percentage of States
Student growth percentiles	14	41%
Value added model	7	21%
Categorical model (value table)	4	12%
Gain score model with vertical scale	3	9%
Improvement	2	6%
Gain score model using *z*-scores	1	3%
Still deciding	3	9%

Student growth percentiles (SGP) were by far the most popular choice for growth models (Betebenner, 2009). Given some early concerns in peer review that a normative model was not appropriate, many states added a criterion component called adequate growth percentiles (AGP). Seven states chose a type of value-added model, a regression approach that is intended to isolate the effect of classroom instruction. Four states used a categorical model, also known as a value table (Hill, 2006), which assigns specific values for movement from one performance level to another across years. The values are typically set by identifying policy priorities, such as whether

moving from basic to proficient is worth more or less than moving from below basic to basic. Values are assigned for maintaining a performance level also. Typically, the higher the level the more points received. However, under the NCLB growth pilot, states were not allowed to assign more points to schools with students moving from proficient to advanced. In the waiver process, two states chose that option. For example, the District of Columbia developed a value table that set values for current year proficiency at 100. Values for current year advanced are 110. Students earn the same number of points regardless of starting position. They also earn zero points for maintaining a level below proficiency or for dropping back a level. To value growth, even below proficiency, students can earn up to 80 points for moving from a low below basic score to a high basic score. These types of values indicate where teachers should focus their efforts to achieve the largest gain.

A couple of states also use simple gain scores from a vertical scale, subtracting the prior year score from the current year score. Although the mathematics may be straightforward, determining adequate growth can take several forms. States can determine what a year's worth of growth is and assign values based on actual growth relative to that; they can calculate the growth necessary to reach the proficient cut score within a certain number of years and count students as on track or not; or states can examine the actual growth in a normative context, giving more credit to students who grew more than others.

States without a vertical scale can conduct a similar subtraction exercise by converting annual scores to standardized scores, otherwise known as *z*-scores. Minnesota uses a standardized scale and a regression equation to predict future performance, and students are given different point values for meeting, exceeding, or not meeting expected scores.

Finally, some states use improvement models, typically called "safe harbor" under NCLB. In these models, performance of third-graders in one year is compared to the performance of third-graders in the previous year. Typically targets are set based on reducing the percentage not proficient by a certain amount each year.

The model used can affect which tests can be included. For example, states using a value table can incorporate results from the general assessment, alternate assessments, and English language proficiency assessments. States using VAM typically only examine growth on the general assessments. In some cases, different models are used on different tests or the tests are analyzed separately so that more students can be included in the accountability decisions. In this way, the growth model can have a significant effect on equity. The CPSD felt strongly enough about this issue to write a position paper, stating:

> Until [states] have a growth model that includes the students taking the Alternate Assessment based on Alternate Achievement Standards (AA-AAS), with expectations of growth towards the same annual measureable objective as all other students, they should not be permitted to use student growth as a major component of their accountability or teacher evaluation systems. (2011)

The Consortium for Citizens with Disabilities (CCD) raised similar concerns in their letter to Secretary Duncan, citing the need for the AA-AAS to be included in all accountability calculations and decisions (CCD, 2011).

Combining Measures

States had to determine how to combine status, achievement gaps, growth, and other factors into a school's accountability determination. Some chose a conjunctive model, requiring

schools to meet minimum criteria on each metric; some chose an ordered, disjunctive approach, requiring schools to meet at least one objective; others combined the measures into an index or school grading system; and still others used a combination of ranking schools on several measures and creating decision rules for categorizing them, a type of compensatory system. Table 5 shows the distribution of states selecting different methods of combining measures.

Table 5

Number and Percentage of States Using Various Methods for Combining Measures

Method for Combining Measures	Number of States	Percentage of States
Index	22	65%
Rank order each measure separately	5	15%
NCLB-type conjunctive	3	9%
Other	4	12%

Three states maintained a conjunctive approach, requiring schools to reach a minimum bar on each measure. Other states rank-ordered schools on each metric and made their judgments based on where a school fell on each list. The vast majority combined the measures into a single score or grade for an overall judgment. For example, the District of Columbia ranks its schools on percentage of proficient students, growth table value, and size of the achievement gap. The first two measures are used to determine reward[5] and priority schools[6], while the third identifies focus schools. Michigan creates multiple lists that each rank schools by (a) the percentage of students at proficient or above; (b) average growth *z*-score; (c) composite improvement in performance

[5] A "reward school" is a Title I school which, based on the most recent data available, is a school with the highest absolute performance over a number of years, or one making the most progress in improving the performance of the all students group over a number of years.

[6] A "priority school" is a school which, based on the most recent data available, has been identified as among the lowest performing schools in the state. The total number of priority schools in a state must be at least 5% of the Title I schools in the state.

levels; and, (d) the size of gaps for the lowest 30% group. The lists are used to include and exclude schools in different orders depending on whether they are classifying schools for reward, priority, or focus. That is, the rules can include percentage proficient below X and *z*-score below Y. At least two states chose to rate schools solely on percentage proficient but continue to use a growth measure as a "safe harbor" that allows low-performing schools to avoid being classified as priority schools if they show high growth from the previous year.

By far, the most common approach was to create some type of index. Almost two-thirds of the states used an index, several choosing to set cut scores on the index that created a school grading system (e.g., A–F or 1–5 stars). One simple type of index is exemplified by Kansas, which assigns points for each performance level attained by each student in each subject in the school. The average score is then taken for that school. Kansas uses a scale with the lowest performance level set at zero points and each subsequent level worth 250 points more, up to 1,000 for exemplary. Mississippi follows a similar approach with different point values.

An example of an approach for combining multiple measures into one number can be found in Maryland, which uses different formulas in high school than in middle and elementary school. Note that within different metrics, multiple subjects are combined. Within achievement, mathematics, ELA, and science are each given equal weight and based on the all students group scoring proficient or advanced. For high school, a school's rating is comprised of 40% achievement; 24% gap reduction in achievement, 8% gap reduction in high school graduation rate, and 8% cohort dropout rate for a combined "gap" measure of 40%; and 10% graduation rate, 4% career attainment, and 6% attendance for a combined "college-and career-readinesss" measure of 20%. For middle and elementary schools, a school's rating is comprised of 30% achievement, 30% growth, and 40% gap reduction.

Kentucky follows a similar approach with a different weighting system that values growth equally or more than achievement gaps, the opposite of what Maryland selected. For high school, a school's rating is comprised of 20% achievement, 20% gap reduction in achievement, 20% growth, 20% graduation rate, and 20% college and career readiness. For elementary and middle schools, a school's rating is comprised of 30% achievement, 40% growth, and 30% gap reduction. Nevada comprises school's scores of 40% Nevada growth model, 30% proficiency rates, 20% subpopulation gaps, and 10% other indicators. These states' varying formulas reveal a range of indicator combinations and ways to weight the various components.

Arizona provides an example of a state that started with a 0–200 index and then converted to A–F letter grades. Of the 200 points, 100 come from growth—50 for all students and 50 for the bottom 25th percentile—and 100 come from the percentage of proficient students. Schools can add to those points with bonuses based on graduation rate, dropout rate, and percentage of ELLs reclassified. Penalties can also be applied for these categories. Indiana followed a similar approach by starting with one rating and then adding or subtracting points based on other indicators. They created a zero- to four-point index based on the percentage of proficient students. Points are added or subtracted based on the growth scores for all students, for the bottom 25%, and for the top 75%. Likewise, Oklahoma focuses on growth in the bottom quartile rather than the achievement gaps. Their ratings are comprised of 33% status, 17% growth, 17% growth of the bottom quartile, and 33% "whole school" indicators, including attendance, dropout rate, school culture, and parent engagement. Only 17% of the rating is based on academic performance of subgroups, and they are identified by percentile ranking, not by subgroup. Although subgroup performance is reported out separately, it is not included in the

accountability determination. The equity issues raised by this type of system and the others described will be discussed further in the section on combining indicators.

Opportunities, Threats, and Monitoring Progress

The remainder of this paper is organized into four sections, each focused on a key aspect of the new accountability systems that relate to monitoring—consolidated subgroups, achievement gaps, growth, and combining measures. Each section identifies opportunities and threats presented by these accountability measures, followed by a discussion of possible monitoring and evaluation approaches. These discussions are intended to highlight the potential benefits of the various state approaches and raise some areas of possible concern that will merit monitoring and evaluation.

Although the USED will monitor the implementation of these new systems, the suggestions here are intended to go beyond the requirements to help the field learn which new approaches truly enhance our understanding of how students learn and to better identify which schools need targeted interventions. The USED monitoring has three components: (1) technical assistance, to support states in their work and identify best practices to help support the work of other states; (2) effectiveness, to examine how a state's implementation of ESEA flexibility is improving outcomes for students; and (3) compliance, to ensure alignment with principles of ESEA flexibility, approved flexibility requests, and Title I requirements still in effect. The monitoring and evaluation suggestions offered in this paper focus primarily on effectiveness.

Consolidated Subgroups

As described in the summary section, 16 of the 34 states that were approved for a waiver proposed a combination of student subgroups for accountability purposes. Ten states created a

numeric group, focusing on the lowest 33, 30, or 25% of students. Other states combined student groups that were persistently low-performing, typically into one or more "at-risk" groups. These groups then contained at least some students from the following subgroups: Black, Hispanic, economically disadvantaged, ELLs, and SWD. Many states have argued that the approach of combining smaller subgroups allows them to hold schools accountable for more students and to focus their efforts on all low-performing students, not just those in a specific category.

The consolidated subgroup approach does provide opportunities to include students in accountability decisions in a unique manner that has not been previously tried. There are, however, some concerns about the approach, both from a civil rights standpoint and from a validity and equity perspective. One concern is that the process of combining subgroups to make accountability decisions could obscure the performances of individual subgroups. The challenges of this approach deal with the choice of methods and monitoring effects, as it is relatively straightforward to implement.

Opportunities.

Under NCLB, states had to specify a minimum *n*-size for including a student group in an accountability analysis. The minimum *n*-size ranged from 5 to 50 across states. In smaller schools, many student groups were not large enough to be included in the subgroup analysis; those students were only included in the "all students" group. In addition, requiring schools to meet AMOs for every subgroup resulted in up to 42 goals for large schools. A school could meet the goals on 41 measures but miss the 42nd and end up in the same category as schools that missed 20 goals. Moreover, any time a conjunctive decision rule is applied, the reliability of the decision is then equal to that of the least reliable measure used.

In developing flexibility waiver applications, there was much discussion on how to reduce the conjunctive decision rule. Adopting a consolidated subgroup approach was one method for reducing the number of judgments made and increasing the reliability of each judgment. When the number of conjunctive decisions is decreased and the size of the groups is increased, each individual decision can have a higher reliability. At the same time, more students were included in the subgroup analysis when the groups were combined because consolidating two smaller groups often leads to one group large enough to meet the minimum *n*-size rule.

Another opportunity in using a consolidated subgroup is being able to better identify students in need of targeted interventions. When students are categorized only by demographic, some students may receive interventions even when they do not need them, while students in a higher achieving group who perform poorly many not receive appropriate interventions. Focusing on the bottom quartile, for example, ensures that the lowest performing students receive attention, regardless of their demographic characteristics.

A key benefit of consolidated subgroups is that they minimize over-identification caused by duplicate counts. That is, under NCLB, an economically disadvantaged Hispanic, English language learner could be counted four times—in each of the three subgroups as well as in the all students group. Under the consolidated subgroup approach, the student would only be counted twice—in the all students group and in the lowest 25% or the highest 75%. Or, by a different classification, the student would fall into the "at-risk subgroup." This student would never be counted in more than one subgroup for accountability determinations.

Threats.

The biggest threat to validity in using consolidated subgroups is that decisions will be made on data that may obscure the performance of smaller groups. Several disability advocacy

groups and civil rights groups have voiced concerns with this approach. The NCLD wrote a letter to Secretary Duncan stating that,

> …the formation of new consolidated groups…for purposes of accountability will mask the performance of students with disabilities. While this approach is often defended for its ability to identify more schools with small numbers of poorly performing student subgroups otherwise not reported because of a state's subgroup (minimum "n") size, combining the performance of several student subgroups does nothing to help schools identify how to go about targeting instruction to the students who comprise the group. (2012)

Additionally, the CPSD wrote in a position paper that,

> SEAs should be required to disaggregate by subgroup for accountability, participation and graduation rate, not just for reporting purposes. They should not be permitted to use a group that combines some or all of the subgroups or some percentage of the lowest achieving students or any other grouping of students that minimizes the impact of the separate subgroups for accountability, participation, graduation and reporting purposes. (2012, pp. 1–2)

To counter the threat that the performance of some student groups would be hidden, the USED required states to continue to set AMOs and report results for each individual subgroup listed in ESEA. However, the guidance allows for states to set different goals for each group, as long as they reach 100% proficiency by 2020, or reduce the percentage of students who are not proficient by one-half in six years. A third option allowed states to create their own AMOs. Of the 34 states that received waivers, 28 set AMOs that call for different levels of achievement for different groups of students. For example, Virginia received a lot of press for their initially-

approved[7] AMOs. *The Washington Post* cited evidence of "soft bigotry of low expectations" in the following statistics drawn from the flexibility waiver: "…schools are expected in 2017 to have 78 percent of white students and 89 percent of Asian students pass the math standards of learning, compared to 57 percent for Black students, 65 percent for Hispanic students and 49 percent for special-education students" (Rotherham, 2012). These targets are quite different from the one expectation per subject and grade set under NCLB, but do follow the instructions under the waiver guidance. However, they have little to do with school classifications. The majority of states with approved waivers identify priority schools as those with the lowest percentage of proficient students for all students and focus schools as those with the greatest achievement gap between two groups. Calculations for determining achievement gaps will be discussed in a later section.

Ultimately, the decisions about inclusion and specification of consolidated subgroups reflect value judgments. Which subgroups are combined and how they are combined could have an effect on which student groups are attended to more closely. A significant challenge lies in closely evaluating the use of consolidated subgroups, and the willingness to change calculations based on the findings. Of particular concern is losing the focus on small subgroups when they are combined with larger ones. If two subgroups, one significantly larger than another, perform differently on the assessment, performance of the smaller subgroup may be masked by the larger subgroup. That is, if the larger subgroup shows marked improvement but the smaller subgroup shows no improvement or even a decrease in performance, the consolidated group would most likely show a small improvement on average. This could result in students who need additional supports or interventions not receiving them. This type of outcome needs to be considered in any monitoring and evaluation plan.

[7]As of October 2012, the USED was reevaluating this approach.

Monitoring and evaluating use.

In *Education Week*, Michelle McNeil reported that Cynthia Brown, vice president for education policy at the Washington think tank the Center for American Progress, stated "If we're going to learn the lessons of this new state flexibility, the federal government is going to have to monitor it carefully and do deep analysis. My concern, very frankly, is they don't have enough resources devoted to it." Much of the monitoring planned by the federal government involves the identification of focus and priority schools and the level and effectiveness of interventions applied. However, this paper will focus more on monitoring the statistics that are gathered before identifying focus and priority schools.

The monitoring required for consolidated subgroups is straightforward; the calculations must be done as described in the proposal. Students should be categorized accurately and, in most cases, their performance will only count in one subgroup. Depending on the approach, consolidated subgroup performance will either be included in an index or used in an achievement gap analysis, both of which will be discussed in later sections of this paper. Of more interest is analyzing the overall use of consolidated subgroups.

For evaluation purposes, it will be important to compare the various combinations of student groups with the original ESEA reporting categories. States still use different minimum sample sizes for inclusion in the accountability systems. In the new waiver applications, the minimum *n*-sizes appear to vary from 5 to 30. With no strong reason for a small minimum n-size working for one state but not another, it will be important to compare the differences in inclusion for states using small minimum *n*-sizes with those using consolidated subgroups. For example, Washington, DC lowered the minimum *n*-size from 30 to 10 but stayed with traditional ESEA subgroups for accountability and reporting. In contrast, Massachusetts initially tried to maintain

its minimum *n*-size of 40 and create a consolidated subgroup of all at-risk student groups; they later agreed to drop the minimum *n*-size to 30 in order to keep the consolidated at-risk group. This situation sets up a natural study between states to see which state includes more students in the subgroup analysis and which is more likely to miss low performance of students in specific subgroups. Another variant is represented by Georgia, which uses a minimum *n*-size of 30 and focuses on the lowest quartile of student performance for its consolidated subgroup. Pertinent research questions for evaluating consolidated subgroups include:

- To what degree is the double-counting of students in multiple subgroups contributing to a school being labeled as "needs improvement"?
- Do consolidated subgroups include more students in the subgroup analysis than using traditional ESEA subgroups with a minimum *n*-size of 10?
- Would the schools identified as "needs improvement" change if they used different methods of consolidation? Options include (a) combining all at-risk subgroups; (b) combining all non-White racial/ethnic groups but leaving the others as standalone; (c) combining ELL with SWD but leaving the others as standalone;(d) focusing on the lowest performing 25%; and (e) focusing on the lowest performing 33%.
- Does the use of consolidated subgroups mask the academic performance of individual student groups? That is, to what degree does improved performance of a larger student group compensate for the lower performance of a smaller student group when they are combined?

Achievement Gaps

Reducing achievement gaps is a hallmark of both NCLB and the ESEA flexibility waivers. The chosen method for calculating an achievement gap is very telling regarding the

specific goals of state programs. As described in the summary section, the majority of states focused on the percentage of students at or above proficient as the metric for analyzing achievement gaps. Only four states chose another metric. Colorado compares the gap between the median growth percentile of each subgroup with the AGP for all students. Georgia calculates the gap using the *z*-score scale. Two other states simply subtract the average scale scores of the two groups and then analyze the size of the difference.

Throughout this section the term "focal group" is used to indicate the students for whom equity is the greatest concern; typically, this is the low-performing group. The term "reference group" is a group to which the focal group is being compared. This group is usually the higher achieving group. For example, a common gap comparison is between low-income students and students who are not low-income. In this instance, the low-income students would serve as the focal group and the not low-income students would serve as the reference group.

Opportunities.

The opportunities associated with including the achievement gap in state accountability systems are unambiguous. First, it permits state policymakers and others to directly gauge inequity, regardless of absolute performance. By tracking gaps in the performance of identified groups, policymakers can readily identify equity concerns that merit attention. Second, it allows state and district policymakers the ability to compare schools with similar demographics but different achievement gaps to better hone in on best practices for reducing gaps. Finally, including achievement gaps sends a clear signal to schools about the outcomes that are valued and allows policymakers to identify schools for specific, targeted intervention.

Threats.

Some threats are also fairly unambiguous. For example, safeguards need to be established to ensure that schools are not credited with decreasing the achievement gap when that decrease occurs as a result of the higher performing group showing a decrease in achievement. More investigation should be done regarding the metric used, the comparison group(s) selected, and the varying effects on schools with different demographics.

Metrics.

The first 34 states approved for NCLB waivers chose from three types of metrics: percentage of students proficient, scale score, and growth score. Each metric provides different information about achievement gaps. States selecting percentage proficient as the metric intend to raise minimum achievement but make no claims about the size of the variance in overall achievement between the highest and lowest scoring groups, provided they are all above the proficient bar. States selecting a scale score metric examine achievement gaps at the smallest grain size. They intend to narrow the achievement distribution across all students. States using growth metrics can address unequal rates of progress for the focal and reference groups ignoring starting points, or address minimizing the gap in rate of growth to proficiency, similar to the percentage proficient metric.

Figure 1 displays different types of outcomes that may be desired with respect to improving achievement for low-performing students. The curves labeled Starting point, End A, End B, and End C illustrate a potential distribution of student performance; the horizontal line indicates the proficiency target. Note that almost all students are below proficient at the starting point. All three endpoints would be acceptable for states that selected percentage proficient as the metric. In End A, the shape of the distribution is unchanged, but shifts up so that most

students are proficient. Still, the gaps persist. End B illustrates a closing of the gap where most students attain proficiency, but students remain clustered near the proficiency threshold. In End C, the variance of the distribution increases above the threshold. Only End B would be acceptable for states selecting some type of scale score as the metric. Ends A or C would be acceptable for those focusing on growth gaps, as all groups would show significant growth.

Figure 1. Various Distributions of Student Scores

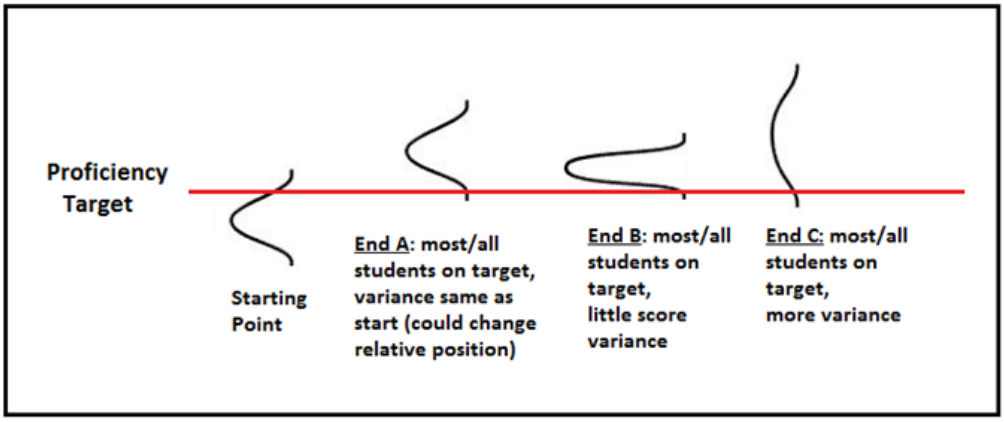

Figure 1. Distributions account for target and score variance across student groups. Adapted from "RILS' Multiple Measures: A personal response," by B. Gong, presented at the Reidy Interactive Lecture Series on September 23, 2011 in Boston, MA. Adapted with permission.

It will be important to monitor both relative and absolute achievement gaps among the various student groups. As End C shows, the percentage proficient gap can decrease while the overall scale score gap increases. Discussion among states about which outcomes are acceptable would be worthwhile.

Comparison group.

State selection of comparison/reference group varies tremendously, which has direct implications on the types of achievement gaps being narrowed. Nonregulatory guidance from the USED instructed states to compare the lowest performing subgroup to the highest performing subgroup. As discussed earlier, this approach could lead to student membership in both groups if, for example, the lowest performing subgroup is SWD and the highest performing subgroup is White students. To avoid overlapping groups, the majority of states proposed other approaches to measuring achievement gaps that maintained the intent while ensuring that each student was counted only once. Furthermore, many states proposed groups in line with the consolidated subgroup that ensured each target group was compared to its opposite. There was concern that the USED instructions could lead to lower priority equity comparisons such as ELLs compared with Asian students. Most states chose to focus not on how subgroups are achieving with respect to each other, but rather how they are achieving with respect to all the other students who are not in that subgroup. Some compared each individual subgroup to its opposite, while others compared the achievement of students in the consolidated subgroup to the achievement of students outside that consolidated subgroup. Still others compared the performance of the target group to the all students group. This statistic, by definition, includes students in both groups. States will need to take precautions to ensure the error term is calculated correctly and appropriate inferences are made.

It will be important for states to monitor the effect of these various decisions over time. Analyzing the achievement gap of consolidated subgroup scores could lead to a decreased emphasis of the gaps between particular groups, particularly when those groups are small. States that followed a lowest percentile approach to identify the target students used various approaches

to select reference groups. For example, one state focusing on the bottom 25% compares the performance of students in that group to the performance of students in the other 75%, while another state uses the top 25% as the comparison group. This distinction could change the outcomes of which schools have the largest achievement gap and are consequently identified as focus schools. Theoretically, the degree of variance and skewness of each school's performance distribution could result in different schools being identified as having the largest gap if the reference group was the top 25%, the top 50%, or the top 75%. Some of these analyses could be done now through various simulations, while others will need to be monitored over time as the changes are implemented.

Monitoring and evaluating use.

Although there is some justifiable variability in the mechanisms used to measure achievement gaps, some outcomes should be avoided. It is important to evaluate gap measures to ensure that the desired policy goals are advanced and that unintended, negative side effects are minimized.

As indicated previously, a significant threat of some measures is that low-achieving focal groups may appear to reflect a small or narrowing gap based only on a low-performing, even regressing, reference group. This threat can be avoided if a common reference group definition is applied for all schools (e.g. the state average or the statewide performance of students who are not economically disadvantaged). The common reference group has significant advantages, but it is not without drawbacks. First, it is important to recognize that use of the common reference group transforms the gap measure to a status measure for the focal group with a "ceiling" applied wherever the common reference group definition is set. Second, achievement gaps can occur above the reference group bar and thus go undetected; this is more problematic if the reference

group bar is relatively low. As a matter of policy, designers may conclude that if the focal group is above the bar (e.g. equal to or greater than the top quartile), this performance is so exceptional that any gaps in this region are not of great concern.

If the system does not use a common reference group measure, it is advisable to examine the relationship between schools evaluated favorably on the gap and the performance of the reference group. For example, in a system based on gaps in percentage of proficient students, it would be useful to plot the gap in percentage proficient with the percentage proficient of the reference group. A strong positive relationship may signal that schools regarded as more favorable are simply those with low overall performance. It may also be useful to examine the distribution of reference group performance (e.g. via box plots) for various levels or ranges of the gap measure. Ideally, the performance of the reference group should have little consequence on the gap. If this is not the case, it may be necessary to adjust the model.

Another type of gap measure is "gap improvement." Some models use improvement in lieu of or in addition to status gap measures. Gap improvement measures offer the advantage that low-performing schools can be credited for progress in reducing gaps over time. However, some applications have the potential for serious shortcomings that should be investigated. As noted previously, declining reference groups can appear to signal progress. Additionally, the magnitude of progress may be distorted by the size of the starting gap. For example, consider the following illustration:

- School A has a gap of five points in year one that was reduced to three points in year two. This two-point reduction indicates a 40% improvement.
- School B has a gap of 60 points in year one that was reduced to 36 points in year two. This 24-point reduction indicates a 40% improvement.

As a matter of policy, it may be desirable to reward the school that demonstrated a much larger magnitude of improvement than the school with minor improvement against a smaller starting gap.

Consider also that this illustration does not address the "location" of the gap on the overall scale. Location refers to the status achievement of the focal and reference groups regardless of progress. That is, are they both relatively high-achieving or low-achieving groups? The focal group of school B might be higher achieving in year two than the focal (or reference group) in school A, which further suggests the accomplishment is very favorable by comparison. On the other hand, school A may be a very high-performing school such that it is exceptional to demonstrate this rate of focal group progress. In fact, some schools may be so high-achieving that it is difficult, even impossible, to show focal group progress because of ceiling effects. Therefore, it is important to understand how gap progress measures interact with gap location. Helpful analyses could include examination of the distribution of gap progress outcomes for ranges of both gap magnitude and gap location. Additionally, if trustworthy growth data are available, such as SGPs, examine the median SGP of the focal groups in the schools rewarded for progress compared with those not rewarded.

Growth

ESEA flexibility waivers have paved the way for states to incorporate growth in state educational accountability systems. In fact, the waivers explicitly included in the criteria for a high-quality assessment a requirement that it "produces student achievement data and student growth data that can be used to determine whether individual students are college and career ready or on track to being college and career ready." This, combined with flexibility for

differentiated accountability systems that reward student progress, led to the inclusion of growth in one form or another in every state application.[8]

Fundamentally, growth simply refers to a measure of student achievement over time. One might regard the flexibility initiative as an opportunity to incorporate "true growth" as restrictions for using growth in state accountability models under NCLB were mitigated. The chief constraint that heretofore stifled innovation was the requirement that growth trajectories must culminate in proficiency in a short amount of time, typically three years or less. States were now free to consider broader approaches to growth, as noted in the summary section.

Opportunities.

State accountability under NCLB was almost entirely based on "proficiency," which creates a simple but coarse dichotomous classification of students based on whether one is above or below a prescribed cut score. In contrast, growth offers the promise of a more accurate depiction of learning gains, essential for a system that values equity. Andrew Ho offers a compelling argument about the shortcomings of systems that rely exclusively on proficiency, warning that it "encourages higher order interpretations about the progress of students that are limiting and often inaccurate" (2008, p. 351). Among the limitations detailed are distortions in trend data (i.e. annual changes in percentage proficient) caused largely by variation in the proficient cut score placement and the characteristics of the distribution. For example, a school with more students near proficient is more likely to demonstrate "gains" than a school with more students well below proficient.

Growth measures present an alternative that can mitigate these concerns. Specifically, in a system that prioritizes equity, incorporating growth can better detect the gains of students who

[8] However, it should be noted that some states did not propose to include growth in school accountability, only in educator evaluation.

are well below the proficient cut score. For example, a value table approach may award points for students moving from level one to level two, or even from categories established within a level. Policymakers may decide to weight progress more heavily to value growth for low-performing students. Many states use the SGP, which evaluates a student's current score, based on the performance of students with similar prior scores (Betebenner, 2009). The resulting growth percentile reflects the student's growth compared with students with a similar academic history, producing a potentially more accurate and meaningful portrayal of performance than proficiency alone.

Additionally, these approaches offer the possibility of establishing more realistic and attainable performance targets for students. Although students far below proficiency may have little chance of reaching proficiency in a short timeframe, thresholds based on growth can provide a more meaningful performance expectation. Such practices may serve to help motivate and monitor the progress of persistently low-performing students. Often states combined norm- and criterion-referenced approaches to establish thresholds for growth. For example, Colorado incorporates AGPs, a criterion measure that indicates the growth rate required to achieve proficiency in three years or less.

Threats.

As discussed in the previous section, growth models offer the promise of helping to define more appropriate and realistic performance targets. Conversely, determining meaningful growth expectations and accounting for sources of error are two threats presented by this approach.

Meaningful growth expectations.

Approaches to setting growth expectations can be characterized broadly as either norm-referenced or criterion-referenced. A norm-referenced approach compares student achievement to a statistically derived expectation, such as the mean performance for students with similar prior achievement. Alternatively, criterion-referenced growth standards establish a specific target outcome. For example, requiring students to grow at rate that will result in achieving a score associated with college and career readiness in a reasonable amount of time is a criterion-referenced approach.

Each approach has advantages and limitations. Setting a norm-referenced expectation is useful for identifying comparably high or low growth. Indeed, it seems intuitively reasonable to describe valued growth as that which is significantly higher than that of similar students. However, a limitation is that some students who grow at very high rates relative to their peers may not achieve proficiency in a reasonable amount of time. A criterion-referenced standard based on reaching proficiency within a certain timeframe resolves this potential "growth to nowhere" problem, but raises a new issue: some students may be so far below standard that even at exceptionally high rates of growth they will not achieve proficiency in a reasonable timeframe. Particularly when growth is used for accountability purposes, this can create a condition where some classes or schools are uniformly disadvantaged. Conversely, very high-performing classes or schools could exhibit little or no growth and remain above performance expectations.

Model error.

Another threat associated with using growth models in accountability is the recognition that multiple sources of error may affect the accuracy and credibility of the model result. In general, growth scores have considerable variability associated with them compared with status

scores. One source is measurement error associated with tests used to compute growth estimates. With traditional criterion-referenced tests used in state accountability systems, it is common to construct the assessment so that most of the information is placed around the cut scores separating one or more performance levels. This is useful when the primary objective is to maximize the accuracy of performance-level classifications. However, such a test is often ill-suited to yield precise outcomes for students scoring at high or low levels. If the assessments are to produce useful information about student growth for high stakes accountability, it must have a "high ceiling" and a "low floor" to measure performance across a broad range. Otherwise, growth estimates will not be sufficiently precise or stable.

Potential variability is also associated with model specifications. Researchers have found that estimated effects are sensitive to model assumptions and specifications (McCaffrey, Koretz, Lockwood, & Hamilton, 2003). In other words, adjustments to model characteristics, such as adding, deleting, or differently defining priors or variables, will very likely produce dissimilar outcomes. Furthermore, missing data can affect the precision and stability of the model and introduce systematic bias in the resulting estimates (National Research Council, 2010). It is generally acknowledged that data are rarely missing at random; rather it is likely that the performance of students with missing or incomplete data differs systematically from those with complete records. For example, mobility rates for economically disadvantaged students, a group which often includes ELLs and Hispanic students as well, are typically higher compared with rates of other students. These concerns are present in virtually any test-based accountability system, but are augmented when the system relies heavily on matched student records required for growth.

Monitoring and evaluating use.

Monitoring and evaluation are critical to ensure that growth measures are producing trustworthy results and being used appropriately to promote the intended goals of the system. The following sections offer important claims that should be investigated in the evaluation process, along with exemplar studies and illustrative evidence. Although not comprehensive, these components are intended to capture the core areas that should be examined to evaluate growth in a school accountability system that is designed to promote equity.

Results are reliable.

Reliability refers to the consistency or stability of a measure. As discussed, the reliability of growth estimates is particularly challenging because of multiple sources of error (National Research Council, 2010). An evaluation plan should certainly include tracking the consistency of estimates across schools, districts, and content areas within each year and across years. Moreover, an analysis of the reliability of growth scores for subgroups (e.g. demographic subgroups or subgroups based on performance) will further reveal the extent to which outcomes are sufficiently stable for all students. Dramatic shifts in results will almost certainly signal a troubling lack of stability that will erode the usefulness and credibility of the growth measure.

Results are valid.

If reliability addresses the extent to which the growth measure provides a consistent answer, validity analyzes the extent to which results are trustworthy and useful for the intended purposes. One advisable set of analyses involves determining the association of growth scores with variables or conditions not intended to be strongly tied to growth. For example, a reasonable number of schools with high poverty rates and schools with traditionally low-performing students should be able to demonstrate higher levels of growth. While a moderate association

with these variables is expected, a strong correlation will call into question the credibility of the results and the extent to which growth is offering information that is not provided through other indicators (e.g. percentage proficient).

Moreover, it is advisable to investigate the extent to which the selected model detects schools judged to be high-performing in the areas aligned with the state's policy values. For example, if the state heavily values academic growth for the lowest achieving students, then the model should be sensitive to detecting progress for students below standard. If there are other trusted indicators, such as schools receiving recognition through existing programs for promoting academic progress, one would expect these schools to earn favorable growth results.

Results discriminate among schools.

Although related to the previous claim, this component addresses the extent to which the overall distribution of outcomes is reasonable. A model in which very few schools receive either unfavorable or commendable results may be out of sync with expectations and the credibility of the results will be suspect. Related to this, if schools of a single type, such as small schools or schools with a homogeneous population of students, all receive similar growth results, this too will discredit the extent to which results are considered trustworthy.

Growth thresholds are appropriate.

Growth expectations should be realistic, but tied to meaningful outcomes. One way to reconcile this is to apply both norm- and criterion-referenced standards, and evaluate the extent to which these expectations produce meaningful outcomes over time. A norm-referenced lens may inspire a state to establish a rubric assigning points to growth levels associated with targets at selected points in a distribution. For instance, growth above the 60th percentile is associated with the highest number of points, growth at or above the 50th percentile receives the second

highest, and so forth. This method provides evidence that the targets are suitably high yet attainable based on prior distributions of student achievement. Accordingly, one aspect of the evaluation plan should involve investigating the distribution of results regularly to ensure the norm-referenced targets continue to meet these criteria.

It is also important to understand the extent to which students and schools receiving favorable growth scores are, in fact, earning or maintaining meaningful outcomes. For example, a straightforward analysis might involve calculating the magnitude of gain for students earning growth scores at or above the top category. It may be particularly illuminating to annually track the percentage of students below proficient who achieve proficiency in various time intervals. If a very small percentage of below proficient students who grow at a rate that is regarded as favorable do not achieve proficiency over time, then it calls into question the suitability of the growth standard to promote and reliably reflect equity outcomes.

Combining Measures and Producing Outcomes

This paper has explored three important accountability indicators related to equity in student outcomes; however, an educational accountability system cannot be evaluated with respect to its component parts alone. A more complete treatment must address how the elements will work together to produce overall outcomes. At least three central questions must be answered to make these decisions:

- How will measures be combined to produce an overall result?
- How will performance expectations be established within and/or across indicators?
- How will results be communicated?

Such design decisions are critical and can have tremendous influence on the extent to which the model functions as intended to promote equity.

Combining measures.

As discussed in the summary section, there are many approaches to combining multiple indicators. The two prominent approaches featured in the ESEA waivers, conjunctive and compensatory, will be discussed in this section. A few states elected to maintain a conjunctive combination rule for all or part of the model. That is, schools must meet minimum standards in multiple categories in order to obtain an overall favorable rating. Obviously, this approach places a strong value on equity by assuring that an overall positive score will not obscure low performance in any key area. On the other hand, conjunctive decisions are typically less reliable because errors accumulate across multiple judgments, meaning that a school classification might be based on the least reliable measure. Moreover, many argue that these policies can systematically disadvantage some schools that have more groups, and thus more opportunities to fail.

Perhaps in reaction to these concerns, most states proposed some type of compensatory approach for combining measures. This refers generally to a method in which higher performance on one measure can offset lower performance on another measure. In many cases, this approach yielded an index, or overall composite score comprised of multiple subscores. A perceived advantage of this method is that it better allows schools with dissimilar patterns of performance to show quality. For example, a school with relatively low achievement test scores may be able to show quality through strong academic growth and/or narrowing achievement gaps. Additionally, classifications based on composites are often more reliable than conjunctive approaches because the overall decision is based on multiple indicators evaluated more holistically. However, indices are not without limitations. Most prominently, summary scores can mask low performance in individual components and portray outcomes as overly optimistic.

For this reason it may be appropriate to add conjunctive decision rules to the system that will serve to protect key policy values. An example of this was the ESEA flexibility requirement that any Title I eligible high school with a graduation rate less than 60% must be classified as a focus or priority school. Relatedly, it is essential to consider appropriate weights for the index that reflect policy values. If a central focus of the system is to privilege equity, substantial influence must be given to the components that reflect real evidence of higher performance for persistently low-achieving students.

The examples detailed in the state summary section showed how some states are using decision rules and weighting to privilege equity, such as by introducing disproportionally higher weights to growth or achievement for identified equity groups, such as students in lowest quartile of achievement.

Performance expectations.

A second factor to consider is the standard for acceptable performance—or "good enough" criteria. Just as interpreting performance on a standards-based test requires the establishment of a cut score to distinguish between performance that is above or below proficient, a method to evaluate performance and classify schools in an accountability system is necessary.

As discussed in the growth section, one frame for establishing performance expectations is to consider a norm-referenced and/or criterion-referenced approach. The criterion-referenced approach should be based on clear policy determinations about what performance is valued. One way to accomplish this using multiple measures is to create profiles of school performance regarded as worthy of selected outcomes. For example, if the state determines policy thresholds for growth, graduation, and proficiency judged to be characteristic of reward schools, the

composite score associated with this profile may be selected as the performance standard. If there are certain ways of achieving this composite score that are deemed outside the boundaries of reward and not aligned with the focus on equity (e.g. very low achievement offset by exceptionally high growth), it may be appropriate to add additional rules ensuring such schools are not classified as reward schools.

The ESEA waiver application required that states identify a certain percentage of Title I schools in the focus and priority categories. Consequently, many states were obliged to include some type of norm-referenced component to inform performance expectations. Often states addressed this by simply ranking schools on overall results or results in individual categories to determine the standards associated with priority and focus categories. Similarly, expectations for reward categories or additional performance thresholds were often informed by considering the distribution of outcomes and identifying thresholds based on a desired percentile.

The two approaches can be blended as well, such as when a state starts with a policy definition for performance and refines the expectation based on information about the distribution of outcomes. Naturally, the criterion-referenced lens helps ensure that key equity outcomes are prioritized, while the norm-referenced approach helps ensure that expectations are "ambitious but achievable."

It is also worth noting that one method of prioritizing equity in accountability is to ensure that standards are set so that classifications produce a manageable number of schools identified for priority support. This allows a state to ensure that adequate resources are directed to the schools most in need, thus optimizing the likelihood that turnaround will be successful. This method, coupled with exit criteria that ensure schools will not move in and out of the priority

support category until convincing evidence of improvement is available, may bolster the likelihood of achieving positive outcomes for schools with the highest equity concerns.

Reporting.

The manner in which results are communicated to the public is another critical aspect of accountability systems that can influence the effectiveness of a system in promoting equity outcomes. Understandably, many stakeholders desire a single, straightforward outcome. This may explain the increase in the number of states assigning a letter grade (i.e. A, B, C, D, or F) to schools, invoking the familiar language of report cards. In other cases, designers will use labels or symbols (e.g. stars) to describe performance. When a single score, grade, or label is assigned to a school, stakeholders receive a succinct message about the school's performance. Another advantage of this approach is that it gives the state control over the way results will be regarded by the public. That is, if multiple indicators are reported without a composite outcome, the media or other groups may determine a manner of reporting (e.g., ranking or averaging) that may not be reflective of good practice.

However, there are some challenges associated with composite scores or labels that are important to address. First, an overall result can mask the important elements that define effectiveness. In so doing, it makes the result less useful or actionable to help stakeholders understand which areas are on track and which need to be improved. Although this can be addressed by reporting both the overall outcome and component results, there are risks with providing information that is too granular. Chiefly, results for smaller groups will be much less reliable than higher level aggregations. Also, it is important to portray the results in a straightforward, easy to understand manner so as to preserve clarity and prevent the "noise" from overwhelming the "signal." .

In general, timely reports should be accessible to stakeholders to ensure that those closest to the students have the information needed to inform instructional decisions. Moreover, the reports should be accompanied by adequate interpretative information. Such information should describe the meaning of and precision of the outcomes and clearly indicate supported uses and interpretations. Even under the waivers, states are required to report out assessment results by every eligible subgroup, which can then be compared to the performance of a consolidated subgroup, if used. Supplemental information may enhance the utility of reports, such as comparative information from similar schools or longitudinal trends.

Another promising practice with respect to reporting is to take advantage of both dynamic reporting technology (e.g. interactive data tables) and data visualization (e.g. graphs and plots). Colorado employs this approach with a system termed SchoolView[9]. In this system, users can access a variety of conventional information, such as summaries of state assessment results, as well as produce and manipulate customized reports.

Monitoring and evaluation of the overall system.

Finally, it is important to consider a plan for monitoring, evaluating, and supporting the overall accountability system. In the best case, the evaluation is ongoing and directly tied to the most important purposes and uses for the system. Many of the evaluation suggestions presented previously are applicable for this section as well. Only the unit of analysis has changed from a component part to the overall outcome. For example, it is advisable to examine the consistency of scores over time and by subgroup, and it is important to determine the association of outcomes with variables or conditions not intended to be strongly tied to results (e.g. poverty, school size). However, there are some additional considerations when dealing with overall model results that merit particular attention.

[9] See http://www.schoolview.org/index.asp for more information including access to dynamic reports.

First, it is important to investigate the influence of sampling error on score consistency. Sampling error refers to fluctuations in school scores that can be unrelated to actual school performance. For example, a school may receive a more favorable accountability determination compared to the previous year because the students enrolled were inherently higher performing, and not because the quality of instruction improved. Naturally, sampling error can work to either support or hinder reported accountability determinations. There are numerous approaches to evaluate the extent of sampling error. One method involves producing multiple sets of results by taking random draws with replacement from the schools to evaluate decision consistency (Hill & DePascale, 2003).

Another useful analysis is to examine discrepancies among indicators to determine if any schools exhibit unusual or problematic profiles. In general, it is expected that indicators will be relatively consistent for schools regarded as high quality. For example, if a school is classified as favorable overall in a compensatory system and it is discovered that this school has high growth and small achievement gaps, but very low status and low graduation rates, this may indicate that the rules for combining indicators need to be revisited. This can be accomplished by evaluating the weights or developing business rules to determine classification of schools with certain score patterns. The main objective is to ensure that that the policy values that prioritize equity are applied so that school profiles that signal the need for support are appropriately classified.

In the best case, ongoing monitoring and analyses should examine the extent to which schools classified as priority and focus are, in fact, improving as a result of interventions. Two pertinent questions for monitoring include: (1) Do schools that exit priority and focus status perform at a satisfactory level for a sustained period of time?, and (2) Does this pattern hold across all indicators for all student groups in all grades and content areas? Such analyses can help

determine where additional supports need to be focused. Moreover, if such analyses reveal that schools exiting support have a high probability of being reclassified as focus or priority in a short amount of time, the exit standard may need to be revisited.

One potential threat that should be monitored is the participation rate of traditionally low-performing subgroups. The flexibility waivers no longer specifically require 95% participation of all subgroups. Although many states maintained this requirement, others used a more nuanced approach to motivate schools to assess all students. For example, Washington, DC, follows up when a school has missed the 95% participation rate two years in a row. Another approach in Arizona decreed that schools can only be listed as an A school with a 95% participation rate; participation rates below 75% automatically make them a D school. Thus, there is a lot of room between 75 and 95% for a school to remain a B school. Participation rates for each student group should be monitored in states with various participation rules to see if there are any negative consequences on equity to easing those requirements.

Finally, feedback from district- and building-level personnel can shed light on the extent to which the system is promoting the intended goals. Do school leaders report that outcomes from the accountability system are clear and actionable? What specific initiatives have been put in place as a result (e.g. curricular revisions, targeted professional development)? How have resources provided to struggling schools helped promote equity? This type of feedback can help illuminate the credibility of the underlying theory of action for how the system promotes desired outcomes.

Conclusion

As stated at the beginning of this paper, educational equity is based on the principles of fairness and justice, and the focus here has been on the aim of promoting improved academic

achievement for low-performing students, particularly those performing below proficiency. The component approaches of consolidating subgroups, calculating achievement gaps, and focusing on growth are all intended to include more students in the school, district, and state educational accountability system and to focus fairly on student outcomes, thus promoting equity. However, in the composite, the indicators could mask performance of individual student groups, particularly groups that are small or less cohesive.

The ESEA flexibility waivers provide a unique opportunity to study multiple approaches to hold schools accountable for raising the academic achievement of all students. The monitoring and evaluation suggestions in this paper are intended to be instructive and present approaches that successfully identify schools in the most need of improvement without allowing others to remain unidentified because of fine distinctions in the indicators or calculations. Such understanding can help to further inform reauthorizations of ESEA.

References

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*, 42–51.

Consortium for Citizens with Disabilities. (2011, December, 22). [Letter to Secretary Arne Duncan]. Retrieved from http://www.c-c-d.org/task_forces/education/CCD%20ESEA%20Waiver%20letter%20to%20Secretary%20Duncan.pdf

Collaboration to Promote Self-Determination. (2011). Position paper on issues related to ESEA flexibility requests. Retrieved August 17, 2012 from http://thecpsd.org/2012/08/17/cpsd-details-concerns-with-esea-flexibility-requests/

Darling-Hammond, L. (2007). The flat earth and education: How America's commitment to equity will determine our future. *Educational Researcher, 36*, 318–334.

Elementary and Secondary Education Act (ESEA) *20* U.S.C. § 6301 *et seq*. (1965).

Fuhrman, S., & Elmore, R. (Eds.). (2004). *Redesigning accountability systems for education.* New York: Teachers College Press.

Gong, B. (2011, September). *RILS' Multiple Measures: A personal response*. Presented at the Reidy Interactive Lecture Series, Boston, MA.

Goertz, M. E. (2005). Implementing the No Child Left Behind Act: Challenges for the states. *Peabody Journal of Education, 80*, 73–89

Hill, R. (2006, April). Using value tables for a school-level accountability system. Paper presented at the National Council on Measurement in Education Annual Conference, San Francisco, CA.

Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice, 22*, 12–20.

Ho, A. D. (2008). The problem with proficiency: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*, 351–360.

Improving America's Schools Act of 1994, Pub. L. No. 103-82, § 108 Stat. 3518 (1994).

Individuals with Disabilities Education Act *20* U.S.C. § 1400 (2004).

Kantor, H. (1991). Education, social reform, and the state: ESEA and federal education policy in the 1960s. *American Journal of Education, 100*, 47–83.

Lee, J. (2004). Multiple facets of inequality in racial and ethnic achievement gaps. *Peabody Journal of Education, 79*, 51–73.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41*, 797–832.

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/MG158

McNeil, M. (2012, October 15). States punch reset button with NCLB waivers. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2012/10/17/08waiver_ep.h32.html?tkn=NSLFJ%2BWQnkqPlIMGUAUBakJda6JiHNTaJZDt&intc=es.

McNeil, M. (2012, August 29). Ed. Dept. gears up to manage NCLB waiver oversight. *Education Week, 32* (2). Pp. 16-17

National Center for Learning Disabilities. (2012, March). State Requests for Flexibility to Improve Student Academic Achievement and Increase the Quality of Instruction. [Letter to Secretary Arne Duncan]. Retrieved August 17, 2012 from www.LD.org

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: Government Printing Office.

National Research Council. (2010). Getting value out of value-added. H. Braun, N. Chudowsky, & J. Koenig (Eds.). Washington, DC: National Academy Press.

No Child Left Behind (NCLB) Act, *20* U.S.C.A. § 6301 *et seq*. (2001).

Rotherham, A. (2012, August 24). Virginia's 'together and unequal' school standards. *The Washington Post*, pp. C1, C4.

Rose, L. C. (2007, September). The sad sage of NCLB. *The Phi Delta Kappan,* 89, 2.