Working Paper: The Waive of the Future: School Accountability in the Waiver Era

Polikoff, Morgan S.[1], McEachin, Andrew[2], Wrabel, Stephani L.[1], & Duque, Matthew[1]

In the decades since *A Nation At Risk*, standards-based accountability has been the most prominent state and federal K-12 education policy culminating with the passage of No Child Left Behind (NCLB), the 2001 reauthorization of the Elementary and Secondary Education Act (ESEA). Despite its promise, NCLB was fraught with problems. In 2011, to help alleviate the impending 100% proficiency deadline, the federal Department of Education began permitting states to apply for waivers to opt out of the NCLB requirements in exchange for implementing their own accountability systems. In this paper, we evaluate the states' wavier applications along four fundamental dimensions of accountability policies: construct validity, reliability, fairness, and transparency. While it is hoped that the federal waiver program will ameliorate the problems with the design and implementation of NCLB, overall the waiver applications provide a mixed bag of progress and regress. We conclude the paper with recommendations for policy-makers on how to improve the quality of the states' waivers.

[2]University of Virginia
Curry School of Education
405 Emmet St. South
Charlottesville, VA 22904

[1]University of Southern California

Updated May 2, 2013.

Center on Education Policy and Workforce Competitiveness
University of Virginia
PO Box 400879
Charlottesville, VA 22904

We would like to thank the participants at the 2013 Association for Education Finance and Policy annual meeting. We would particularly like to thank Randall Reback for his thoughtful feedback on an earlier draft. The project was partially supported by IES Grant R305B100009 to the University of Virginia. The views expressed in the paper are solely those of the authors and any errors are attributable to the authors.

THE WAIVE OF THE FUTURE: SCHOOL ACCOUNTABILITY IN THE WAIVER ERA

Polikoff, Morgan S., McEachin, Andrew, Wrabel, Stephani L., & Duque, Matthew

In the decades since *A Nation At Risk*, standards-based accountability has been the most prominent state and federal K-12 education policy. No Child Left Behind (NCLB), the 2001 reauthorization of the Elementary and Secondary Education Act (ESEA), created the first mandatory national accountability structure that held schools and districts responsible for student achievement. Despite its promise, NCLB was fraught with problems (Balfanz, Legters, West, & Weber, 2006; Ho, 2008; Linn & Haug, 2002; McEachin & Polikoff, 2012; Porter, Linn, & Tremble, 2005). In 2011, to help alleviate the impending 100% proficiency deadline, the federal Department of Education began permitting states to apply for waivers to opt out of the NCLB requirements in exchange for implementing their own accountability systems. Currently, 35 states have applied for, and won, these ESEA waivers.

The theory of action of accountability policies posits that the use of incentives (sanctions/rewards) will motivate educators to align their instruction and behaviors with a set of predetermined standards and outcomes (Smith & O'Day, 1990; Figlio & Ladd, 2008). However, accountability policies that hold schools accountable for student outcomes rely on a number of important implicit assumptions that impact on the opportunity for these policies to trigger the desired changes in student outcomes. When these assumptions are violated, accountability policies can lead to unintended consequences that are potentially harmful to students' success.

While there are a number of important aspects of accountability policies, we focus on four fundamental assumptions of accountability policies using measurement research: construct validity, reliability, fairness, and transparency (AERA, APA, & NCME, 1999; Baker & Linn, 2004; Kane & Staiger, 2002; Linn, 2000 & 2002). In what follows, we define each of these assumptions and illustrate how they have played out under NCLB's accountability system. Due to the dynamic and

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

multifaceted nature of the schooling process, it is unlikely that accountability policies are able to capture all of the important aspects of an educator's role. Therefore, it is likely that any accountability policy will violate at least one of these assumptions in some degree. The question then becomes to what extent the policy violates the assumptions, and what the potential unintended consequences of those violation are.

In this paper, we evaluate the extent to which the accepted ESEA waivers violate these four assumptions, using relevant research to guide our assessment. It is hoped that by fleshing out the potential violations of these four assumptions early in the ESEA waiver process policy-makers and educators will have a better understanding of how the design of accountability systems has a direct impact on the effectiveness of these policies.

Background

Standards-based Reform

The rise of public education accountability policies owes its beginnings to the larger systemic reform, or standards-based reform, movement of the 1980s and 1990s (Smith & O'Day 1990). Standards-based or systemic reform systems contain at least the following six components (Hamilton, Stecher, & Yuan, 2008): 1) clear academic expectations for students in the form of curricular frameworks; 2) alignment of the key elements of the educational system; 3) the use of assessments to measure student and school outcomes; 4) decentralization of resource allocation, curriculum and instruction to schools; 5) technical assistance or support from states and districts to low-performing schools; and 6) the use of accountability policies that reward or sanction schools based on measured school performance. School accountability policies are therefore viewed as a subset of the larger systemic reform movement, not as a standalone reform (O'Day and Smith 1993;

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Smith and O'Day 1991). Only after a state put in place the first five SBR components were schools and districts to be held accountable for their students' achievement.

No Child Left Behind

The most recent federal implementation of educational accountability is NCLB. In 2001, Congress reauthorized ESEA, requiring the design of school accountability systems in order for states to receive federal funding. Accountability under NCLB requires schools to meet progressive proficiency targets each year, culminating in the expectation that 100% of students are grade-level proficient by 2014. Annual measurable objectives (AMOs) are the annual interim proficiency targets schools must meet in both mathematics and English/language arts (ELA). Schools are accountable for performance disaggregated by numerically significant subgroups based on racial/ethnic, disability, socioeconomic, and English language proficiency status. The federal government left to states the flexibility to define proficiency, determine the progression rate of AMOs, and establish the minimum number of students necessary to consider a subgroup significant. These decisions had important implications for the number of schools identified as failing across states (Balfanz et al, 2007; Porter, Linn, & Tremble, 2005).

In addition to meeting minimum proficiency targets, schools are also held accountable for a 95 percent participation rate for each subgroup and the overall school. Finally, graduation rates must be included in high school performance calculations and an additional measure of performance must be identified by the state for elementary and middle schools. Schools meeting all of their AMOs are classified as making adequate yearly progress (AYP); thus, missing one AMO in a year results in AYP failure. While there are several alternative methods for making AMOs, some of which are widely used (Polikoff & Wrabel, 2013), large majorities of schools in many states are now failing to

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

demonstrate AYP (Balfanz et al, 2007; Sims, 2013). Schools that fail AYP and that receive Title I funding are subject to increasing sanctions for failing AYP.

Literature on Accountability Systems

Broadly, there are two primary streams of economic theory that support the use of accountability in education. The first is principal agent theory (Holmstrom and Costa 1986; Milgrom 1988; Milgrom and Roberts 1988), which suggests that the incentives created through accountability systems can help direct educators' efforts toward those behaviors most important for improving student outcomes (Acemoglu, Kremer, and Mian 2008; Kremer and Sarychev 2000; Stein 1988; Prendergast 1999). The second is the experiential goods literature (Shapiro 1983), which argues that the infusion of quality information can help educational consumers (e.g., parents, students) make better choices from among educational options (Figlio and Loeb 2011; Jacobsen, Snyder, and Saultz 2012; Rothstein, Jacobsen, and Wilder 2008; Charbonneau and Van Ryzin 2011). While a full review of these strands of literature is outside the scope of this paper, the benefits of accountability under each theory are heavily dependent on the specifics of the type and quality of the information provided to educators and consumers. In the next section, we use the measurement literature to describe four key features that define high quality data necessary to help parents and others make important educational decisions.

Construct validity. In our context, construct validity is the set of defensible inferences that can be established from a subset of performance measurements (Cronbach & Meehl, 1955; Crocker & Algina, 2008). For example, an accountability policy has construct validity if the performance measures adequately cover the latent (or unobserved) set of desired student outcomes, and if the inferences made on the basis of those performance measures are appropriate. Accountability policies implemented to date typically rely on objective measures of school and district performance. It is

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

assumed that while goals such as active citizenship, ethics, and critical thinking are important but left unmeasured, holding schools accountable for aggregate test scores on math and ELA exams closely proxies these unmeasured goals (Rothstein, Jacobsen, & Wilder 2008).

No Child Left Behind provides a good example of the construct validity problems that can arise within accountability policies. First, NCLB's use of a status measure of achievement, rather than growth, does not account for schools' contributions to student learning (Heck 2006; Kim and Sunderman 2005; Krieg and Storer 2006; Weiss and May 2012). Second, NCLB's use of proficiency rates makes it difficult to measure progress over time since changes in proficiency rates are unstable and measured with error (Ho 2008; Linn 2004; Kane and Staiger 2002). Third, although NCLB does allow states to use growth-to-proficiency models, these models do not meaningfully account for school improvement (Ho, Lewis, and Farris 2009; Weiss and May 2012; Polikoff and Wrabel, 2013). Fourth, NCLB's focus on only mathematics and ELA proficiency undoubtedly falls short of capturing all the important outcomes of schools.

Reliability. We define reliability as the consistency of a performance classification either between multiple measures at the same time (e.g., proficiency versus value-added) or between the same measure at multiple time points (e.g., year-to-year stability) (Crocker & Aligna, 2008). While we could use a broader definition that concerns the absolute score of schools on a given measure (e.g., the spearman-rank correlation between proficiency rates and value-added measures), this work is more concerned with the classification in performance categories.

A number of factors can affect the reliability of the performance measures; the two most important of these are the type and level of information and the number of years. Under NCLB, schools are primarily accountable for the proportion of students that score proficient. This measure is highly reliable – indeed, schools falling below NCLB proficiency cuts often find themselves

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

unable to improve enough to escape sanctions. However, because NCLB proficiency targets are rapidly increasing, schools are perhaps more accurately accountable for changes in proficiency rates. These changes in yearly proficiency rates are very noisy due to measurement error and sampling variation (Kane & Staiger, 2002).

The second important factor regarding reliability is the number of years used to generate the performance measure. Under NCLB, schools are held accountable for their students' average raw test scores, or changes in their test scores between two adjacent school years. Random fluctuations in students' test scores (e.g., a particularly noisy testing environment in a given year) decrease the ability to generate reliable estimates of schools' performance. Increasing the number of years used to generate a school performance measure can reduce the potential bias due to random fluctuations (McEachin & Polikoff, 2012).

Fairness. We define fairness in accountability systems as the level of disparate performance classifications of schools according to their demographics (Camilli, 2006). As noted by Camilli (2006), "The purpose of a fairness investigation is to sort out whether the reasons for group differences are due to factors beyond the scope of the test [performance measure] (such as opportunity to learn) or artifactual" (p. 225). Another way of thinking about the fairness of an accountability system is to consider the reference population for a given school. The population can either be absolute (i.e., all schools) or conditional (e.g., schools that share similar student populations) (Barvley & Neal, 2012; Elhert et al, 2013) – the latter would be fairer.

A fair accountability system would be one that holds schools accountable for only the portion of student achievement they can actually control. A few pre-NCLB accountability systems used statistical adjustments to remove the variance in students' test scores that was unrelated to school-controlled factors (e.g., race/ethnicity, socio-economic status) (Clotfelter and Ladd 1996).

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

However, the current NCLB system holds schools accountable solely for the percent of students that reach a state-defined proficiency threshold. Thus, research clearly shows that schools with more significant subgroups (i.e., larger, more diverse schools) are more likely to fail AYP (Balfanz et al. 2007; Krieg & Storer 2006), as are schools with lower initial achievement (Riddle and Kober 2011). Clearly, NCLB's AYP is not a fair measure given its bias against diverse and previously low-achieving schools. Indeed, unless an accountability policy makes specific provisions to control for non-school related factors of student achievement, it is likely the system will exhibit some degree of unfairness (Balfanz et al. 2007; Krieg & Storer 2006).

Transparency. We define transparency as the level to which the performance goal-setting process is clearly documented and the performance measures are clearly understandable (AERA, APA, & NCME, 1999). Baker and Linn (2004) provide several relevant suggestions regarding transparency: 1) If indices or weighted averages of multiple performance measures are used to hold schools accountable, then the specific weights should be coherently and explicitly stated; 2) In the design of accountability systems, the expectations of all parties (e.g., students, teachers, parents) should be made public and understandable; 3) If student assessments are used to make inferences about school quality, then data should be provided to explicate that the assessments are sensitive to instructional quality and student effort; 4) If schools are classified based on student assessments, then information about the error rates and quality of the assessments should be make public; and 5) Yearly reports provided to stakeholders should promote the valid interpretation of the results from students' assessment and school classifications.

No Child Left Behind's accountability measures fare reasonably well in terms of transparency. The use of proficiency rates is more straightforward than other potential achievement measures. However, the lack of a common meaning for proficiency across states (National Center

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

for Educational Statistics, 2007) is troubling and less transparent than is ideal. Furthermore, there are numerous alternative methods to make AYP other than meeting all the AMOs; these are not well documented or understood, but they account for an increasingly large share of schools passing AYP (Polikoff & Wrabel, 2013).

ESEA Flexibility

Flexbility Policy

The provisions of NCLB remain in effect because the ESEA was not reauthorized in 2007 as scheduled. However, the USDOE has recognized that parts of the law have become barriers to implementation of innovative education reforms (USDOE, 2011). While waiting for reauthorization, Secretary of Education Arne Duncan offered states the opportunity to request flexibility from certain NCLB mandates in exchange for a state's pursuit of rigorous and comprehensive plans to reduce achievement gaps, improve instruction, and advance educational outcomes. As of May 2013, all states but Nebraska and Montana had submitted flexibility requests (or "waivers"). Thirty-five requests have been approved, with the rest under review or rejected.

The USDOE has identified four waiver principles. The principle relevant to school accountability is called "differentiated recognition, accountability and support." The first requirement under this principle is to identify which subject areas will be included in testing and accountability. Second, states must outline their new AMO structure, either creating their own plan or choosing between two given options: AMOs that increase in annual equal increments and result in 100 percent proficiency by 2019-2020 *or* reduce by half, within six years, the percentage of students in the "all students" group and in each subgroup who are not proficient. Third, states determine how subgroups will be identified using either the NCLB subgroups or creating their own. Fourth, states must outline how each they plan to evaluate school performance using the new

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

AMOs and indicators of school performance. Flexibility guidelines require a measure of student growth be included in the calculation of school performance to reduce the misidentification of progressing schools and more appropriately identify and support chronically low performing schools (USDOE, 2010).

Finally, the identification and consequences for reward, focus, and priority schools must be outlined. The highest-performing schools or highest-progress schools are identified under the reward classification. Schools in both categories may not have large achievement gaps between significant subgroups. Focus schools are defined as Title I schools that contribute to a state's achievement gap. States must identify 10% of Title I schools with the largest within school gaps between high and low achieving subgroups (or between graduation rates in high schools). Title I high schools with graduation rates below 60% over a number of years must also be identified as focus schools. Priority schools are defined as the state's lowest performing schools. The total number of priority schools in a state must be at least 5% of the Title I schools in the state. A priority school may be a school among the lowest five percent of Title I schools in the State based on the achievement of all students, a Title I-participating or Title I-eligible high school with a graduation rate less than 60 percent over a number of years, or a Tier I or Tier II school under the School Improvement Grant (SIG) program that is using SIG funds to implement a school intervention model. For this analysis, we focus on the identification of priority and focus schools for accountability, because the AMOs discussed above are not necessarily used by the states for accountability purposes. Furthermore, non-consequential accountability gives schools much less pressure to improve (CITE Dee).

Submitted and Approved Waivers

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

As of May 2013[1], 48 states, the District of Columbia, Puerto Rico, and the Bureau of Indian Education have submitted requests for ESEA flexibility, as seen in Table 1. Of those waiver requests, 35 have been approved, 15 are under review, and one was rejected. Six unapproved states received permission to modify or maintain their proficiency targets for the 2012-2013 academic year. Only Montana and Nebraska have not submitted applications. We focus only on approved flexibility requests for this paper.

Methods

In order to compare the requirements of each state's newly designed accountability system, we analyzed each approved flexibility request in a three phase process. Two research team members reviewed each waiver through the first two phases, and each reviewer analyzed each waiver during either the first or second phase of analysis. At any point if two reviewers disagreed on how to interpret a specific aspect of a wavier application, a third reviewer was consulted on the appropriate classification.

For the first phase, the waiver applications were read in their entirety. Waivers were outlined and condensed according to the four waiver principles. The second phase utilized the condensed waiver outlines to code the specific accountability designs. Design features coded included such things as subgroup size, subjects tested, components and weights of composite indices, and growth measures. Finally, we applied the four measurement dimensions of accountability policies to describe each waiver application.

Below, we summarize our results pertaining to the four dimensions mentioned above. In each case, we provide counts of the number of states meeting certain criteria. Where applicable, we

[1] All information taken from (USDOE, ESEA Flexibility http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html).

also provide examples describing particular features in greater detail. We exclude Washington state from all analyses because they did not describe the index to be used to identify priority or focus schools in their application. Thus, there are 34 applications represented in this analysis. It is important to note that the vast majority of states identify priority and focus schools in multiple ways. For instance, half of the priority schools in a state might be based on a composite index, while the other half might be based on having graduation rates below a certain threshold. The numbers that follow often add up to more than the 34 states we analyzed.

Results

Construct Validity

There are several ways in which the identification of priority and focus schools in the waiver plans is, in aggregate, superior in terms of construct validity to the way schools were identified under AYP. For one, many of the waiver accountability systems include non test-based measures in the identification of low-performing schools. For priority schools, 19 states identify schools using a measure of graduation rates; here, the most common rule is that schools with less than 60% graduation are identified. Separately, 22 states use a composite index of school performance (e.g., an A to F grade based on a combination of performance measures) to identify priority schools. Of these 22 indices, 17 include graduation rates and 11 include other non test-based measures. The most common additional measure is some college/career ready indicator (nine of the 11), but states also include attendance, test participation rates, teacher/principal effectiveness, school climate, and opportunity to learn measures. Only Wisconsin and Arkansas use test scores alone in the identification of priority schools.

Focus schools are more often identified by state student achievement test data. Twenty-seven states use either proficiency gaps (either within school or between a subgroup and a statewide

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

average or target) or subgroup proficiency rates to identify focus schools. Still, the majority of states use at least one non-test-based measure for identifying focus schools. Graduation rates or subgroup graduation rates are used to identify focus schools in 16 states. Eight states use their composite index to identify focus schools; all eight have graduation rates in their composite indices, and six have other non-state-test measures.

A second way in which the identification of focus and priority schools is superior to AYP in construct validity is that many of the priority and focus classifications use test-based measures other than proficiency rates. While none of the waivers uses student achievement growth as a separate criterion for identifying schools, 16 states using a composite index have growth as a component of the index. The weight on the growth measure in the composite index ranges from 75% for Idaho elementary schools to 14% for Kentucky high schools. Some states do not assign explicit weights to growth models, but rather count students as proficient for the purposes of calculating proficiency rates if they have passed their growth target. The vast majority of these growth measures use student growth percentiles (Betebenner, 2011) or some variant thereof. While these measures may not be ideally suited for identifying the effect of schools on students' learning (a true value-added model is probably superior (CITE)), these measures are much closer to identifying schools' contributions to student learning than AYP's percent proficient. Even in states where achievement levels remain an important part of the identification system, 11 are moving from proficiency rates to a system that allocates points along the full distribution. This is a stronger approach, in that it does not dichotomize the continuous achievement distribution. Finally, for focus classifications, two states account for subgroup-specific growth rates and nine states use a proficiency gaps measure. These types of measures directly target the goal of focus schools – identifying and reducing achievement gaps.

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

While these are promising signs in terms of construct validity, the fact is that almost anything would have been an improvement over AYP, which was based only on proficiency rates in math and ELA. There are two main shortcomings with the waiver applications in terms of construct validity. For one, the majority of the states are still using just mathematics and ELA to determine priority and focus schools (21). And 17 states are actually removing science testing altogether, which is a step backwards from NCLB (where science was tested but not included in accountability). Of the 13 that are using other subjects, all 13 are testing science, six are testing writing, five are testing history or social studies, and one is testing other subjects. Still, many of these states are testing these subjects in only a few grades, and most are still giving the preponderance of the weight to mathematics and ELA. Thus, there will still be strong incentives to focus on mathematics and ELA in the majority of states.

Second, while some states include creative non-test measures in their indices, these rarely account for a substantial proportion of the total and are mainly for high schools only. Again, this incentivizes educators to focus on the test scores, which almost always make up 70% or more of the total index score. Furthermore, a remarkable number of states continue to emphasize proficiency rates, despite all we know about the poor quality of this measure as an indicator of school performance. One way in which proficiency rates remain a primary metric is in the identification of SIG schools as priority schools, which is the case in all but one state. Also, using raw graduation rates has similar shortcomings to using proficiency rates, such as strong correlation with student demographics. Overall, our conclusion is that the construct validity of most states' methods to identify schools is better than under AYP, but there are still many obvious shortcomings of these systems.

Reliability

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Evaluating the reliability of priority and focus classifications is more difficult, because the systems have not been implemented. However, we know that AYP was highly reliable in identifying low performing schools, because school-level proficiency rates do not change much from year to year (McEachin & Polikoff, 2012). On the one hand, most of the waiver states plan to identify their focus and priority schools only every two to three years. By definition, this will mean less year-to-year fluctuation in these classifications than would otherwise be the case.

On the other hand, there are several reasons to think that the reliability of the priority and focus classifications will be lower than that of AYP. The first is that the priority and focus classifications are based on a fixed percent of schools. This kind of norm-referenced cut score results in decreased reliability, because there is substantial measurement error and imprecision in these performance indices (Kane & Staiger, 2002; Ho, 2008). Thus, there is not likely to be meaningful differences in performance between schools in the 4th percentile and those in 6th percentile, and yet under the waivers these schools would be treated differently. While they might be classified as a focus school if they just miss the cut for priority, they might not given that most states use different measures to identify focus and priority.

A second reason there will be decreased reliability is that many states are using growth models in their indices. Even school-level measures of student growth are fairly unstable from year to year (CITES). While states could use multiple years in their growth measures, only five states chose to do so. Given the unreliability of these growth measures, the composite indices will be less stable than the percent proficient index used under AYP. In short, while using growth data enhances the construct validity of the performance measure, it is likely to decrease the reliability of the measure.

Fairness

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

The approved waiver plans are likely to be substantially unfair to schools serving large proportions of historically low-performing subgroups. However, the fairness will be better than under AYP. As mentioned above, there remains a heavy reliance on status-based measures of achievement. Even in the states that use indices to identify priority schools, a) proficiency rates are a substantial proportion of the index in every state, and b) these states also identify SIG schools (which were themselves identified by low proficiency rates) as priority schools. Proficiency rates also factor in the identification of focus schools: 19 states identify focus schools using either subgroup proficiency or a subgroup index based on proficiency rates, and an additional seven use their composite indices, which also have heavy status components. Subgroup proficiency-based measures target schools serving diverse populations, and particularly those serving large proportions of students with disabilities (McEachin & Polikoff, 2012). To be sure, some states have established annual measurable objectives that differ for each subgroup – in some cases resulting in local controversy (McNeil, 2012). These subgroup-specific AMOs would be fairer than targets that are the same for all subgroups, but none of these subgroup-specific AMOs are included in priority or focus classifications.

Some states are also employing one of two types of proficiency gap measures for identifying focus schools. Within-school proficiency gaps compare the magnitude of the largest gaps between two subgroups within a school; these gap measures indeed target schools with large within school performance gaps. They may still be unfair to diverse schools, but the degree of unfairness is likely not as large as for simple status measures. The second type of proficiency gap measure compares the performance of a subgroup in a school to a state average or other statewide target. Though these are called "gap" measures, they are actually subgroup status measures.

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

While diverse schools will be more likely to be classified as failing under most any accountability system save one that explicitly controls for student demographics, there are some ways in which the approved waiver applications will decrease the diversity penalty. One way is in states that use so-called "super subgroups" rather than NCLB subgroups. Super subgroups generally take two forms. One is a subgroup consisting of a combination of traditionally low-performing subgroups. For instance, Mississippi's super subgroup includes all students in any traditionally low-performing subgroup. The other is a subgroup consisting of the lowest performing students in a school. For instance, Michigan's composite index includes a gap measure that compares the achievement of the top 30% to that of the bottom 30% in each school. Both of these approaches will tend to reduce the diversity penalty currently experienced by schools, though the latter type will likely reduce the diversity penalty more since it is agnostic to student demographics.

Composite indices or classification schemes incorporating growth models will also tend to be fairer than those based more heavily on status measures, because the correlations of growth measures with student characteristics are smaller (Elhert et al, 2013). However, these correlations will not be zero, and some have expressed concerns that growth models might again be biased by nonrandom sorting of students across schools (Elhert et al, 2013). One way to largely control this problem would be to explicitly control for student demographics in a value-added model (CITE). However, no states using growth models have chosen to do so. Quite the contrary, only one state (New Mexico) has proposed using a value-added model in its composite index, and this model does not include any student demographics. All 16 other states using growth models use models such as student growth percentiles or contingency table approaches that account only for students' prior achievement. These approaches, while fairer than status approaches, may still be unfair to schools serving more students from historically disadvantaged subgroups.

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Transparency

While NCLB's AYP system was unfair and had weak content validity, the use of percent proficient was reasonably transparent. On the surface, the new grading systems in place in most states are even more transparent. Many states use either an A-F or point-based index system, which condenses multiple measures of school performance into a single aggregate. These indices, because of their familiar form, should be fairly interpretable by educators and the public.

However, there are several problems that limit the transparency of these indices. Perhaps the most glaring is that many states have composite indices but do not use them to identify either priority or focus schools. Indeed, among states with a composite index, 10 states do not use it to identify priority schools and 16 states do not use it to identify focus schools. There are another nine states that use either a modification of their index or only some portion of their index (e.g., one subscale) to identify focus schools. For instance, Minnesota recalculates their regular index methodology but applied only to the historically low-performing subgroups to identify focus schools. In all of these cases, the schools with the lowest composite index will not necessarily be the ones identified as priority or focus, sending potentially confusing messages to educators and the public.

Less clear still, in some of these states there is a third method of measuring performance for the AMOs. In Nevada, for example, there is a composite index based on student achievement levels, growth, subgroup growth, graduation rates, and other indicators. But for priority school identification the state excludes the subgroup growth measure from the index, and the AMOs are based only on percent proficient. Again, this kind of system may send unclear messages to teachers and parents, because the priority and focus schools may not be the lowest performing schools on either the index or the AMOs.

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Another way in which the transparency of the measures is unclear is that many state indices apply confusing, complicated, or seemingly arbitrary weights to unrelated measures to arrive at a composite score. For instance, South Dakota's School Performance Index is a 100-point scale. For elementary schools, 25% of the grade is based on ELA and math proficiency rates, 25% on a growth model, 20% on attendance, 20% on principal and teacher effectiveness, and 10% on school climate. For high schools, 25% is based on proficiency rates, 25% on graduation rates, 20% on college and career readiness, 20% on principal and teacher effectiveness, and 10% on school climate. In each case, the points are simply allocated in proportion to the raw values – for instance, a school with 70% of educators rated "proficient" or "distinguished" would receive 14 points for that component. Thus, while the 100 point index is conceptually transparent, it is not clear whether a school scoring, for example, 80, is doing an effective job. Furthermore, it is not immediately apparent from looking at the index score what a school might do to improve its score.

Yet another challenge in some state indices is that the calculation of the subcomponents. One common challenge is in the use of contingency tables to transform continuous variables for the purposes of inclusion in the index. For instance in Idaho, each of the subscales on the composite index is first converted to five- or ten-point sub-indices, which are then weighted and added. As an example, proficiency rates are converted as follows: ≤ 40% = 1 point, 41-64% = 2, 65%-83% = 3, 84-94% = 4, and 95%+ = 5. Median growth percentile scores have their own contingency table (in fact, two – one each for schools that do and do not meet their growth targets). These kinds of approaches are conceptually unclear, and states rarely offer rationales for their use. These contingency table approaches appear to suffer from the potential "bubble-kid" problems associated with NCLB's AYP system (Booher-Jennings, 2005). With several contingency tables in each index,

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

schools could merely determine on which index they are closest to the next point bump and target their efforts at that component.

Finally, some of the calculations states are proposing to use in their indices or to identify priority or focus schools are simply unclear, even after repeated readings by multiple experts. In these cases, there is no question that parents and teachers will have difficulty understanding the index. Perhaps the most challenging index to understand is Missouri's. Missouri has a 40-point index (60 points for high schools) and uses contingency tables similar to Idaho's. A total of 32 points are given for achievement status, 16 per subject. For each subject, there are 28 possible points, 16 for performance levels and 12 for either progress or growth, whichever is higher. Progress is based on a percentage reduction in a gap between the performance index and a target (akin to reducing the percent below proficient). Growth is based on a student-level growth model that is not described in the application. Even if a school earns 28 points for a subject (by rating in the highest range in both status and growth or progress) that school would get only 16 points. These same measures are replicated for the gap group of students in historically low-performing subgroups, but at one-fourth the weight. In short, this index is challenging to understand, uses contingency tables with no rationale, and offers no incentive to improve for any school earning maximum points on performance levels.

Discussion

As the deadline of 100% proficiency by 2013-14 looms, there has been increasing pressure on the federal government to update and/or revise the ESEA law. Given the inaction with reauthorizing ESEA by the Congress, the USDOE implemented a waiver program in 2012 to help alleviate the burdens of NCLB on states. As we have discussed in this paper, the waiver program provides states the opportunity to implement their own accountability systems, often substantially

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

reducing the number of schools identified as failing in a given year. As we have also shown, the waivers provide a mixed bag of improvements over, and duplications of, the problems associated with NCLB.

In many of the waiver applications, states have strengthened the construct validity of their accountability systems by using a combination of growth and proficiency and moving to larger, more stable, super subgroups. It is hoped that these changes will capture more of the multidimensional nature of the schooling process, and increase the alignment between incentives and outcomes in accountability policies.

In most states, however, it appears that many of the problems associated with NCLB will be duplicated. For example, while states were allowed to include additional tested subjects in accountability, half of the states actually took away a tested subject (science), and nearly every system includes a heavy focus on math and ELA. Also, the heavy reliance on proficiency rates to identify priority and focus schools will continue to over-identify schools serving students from historically disadvantaged groups. While states were given the opportunity to increase the reliability of their measures by using more than one year of data, most states chose to only use one year's data when identifying failing schools. This decision will likely reduce the reliability compared to the NCLB system.

The waivers also offered an opportunity for states to incorporate growth measures into their accountability systems. However, the type and use of growth measures across the waivers dramatically. The measures range from NCLB-like changes in school proficiency rates to the newer student-level Student Growth Percentiles. In all cases, states are not controlling for student demographics; therefore, schools' performance will likely be biased by student demographics.

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Furthermore, a number of states that are implementing growth measures are not using them in the identification of priority and focus schools.

As noted in the introduction, it is likely that any accountability system will violate one of the four important measurement aspects. The waiver applications perfectly illustrate these tensions. While many of the waivers provided positive changes over the NCLB system, the net change is harder to summarize. We close with a few policy recommendations that mitigate the unintended consequences of the current waiver applications, and that are not difficult to implement.

Policy Recommendations

The first and most important policy recommendation is to incorporate the lessons learned from NCLB into the implementation of the waiver applications. A number of the waiver applications implement policies that are known to pose specific problems (e.g., the heavy use of proficiency rates, the focus on two subjects only).

States should create more refined comparison groups for schools by conditioning on student demographics in the construction of school performance measures. By not including student demographics in any of the performance measures, the system expects the same performance from all schools, regardless of their students' characteristics. While this comparison has certain strengths in terms of equity, it ignores what we know about student learning trajectories, and it holds schools accountable for factors they cannot control. This unfairness contributes to unintended consequences such as teachers preferring to work in schools serving more affluent children.

For subgroup specific measures, states should move away from within-state performance gaps to within-school or within-district. This would change the focus away from low performing subgroups (already a major focus) to reducing the gap within a school or district. A within school or

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

district system would send a clear message that all students within a school deserve a high quality education.

States should use multiple years of data to improve the reliability of school performance measures. By now, most states have the ability to use multiple years of data in the construction of school performance measures.

States should move away from an arbitrary norm-referenced criterion for identifying low performing schools. Although setting the bar at the bottom 5% or 10% creates a system with a more manageable sample size, it also adds noise to the system. By design, the use of these cutoffs sends the message that 10% is failing but 11% is not, even though there may not be meaningful differences between these schools. Instead, states would likely benefit from a clear operational definition of a low performing school that is based on a set of performance criteria.

States should conduct short-term analyses of the implementation of their waiver systems. The problems of NCLB were well known shortly into the implementation, yet little was done to mitigate the unintended consequences.

None of these recommendations, on their own, will solve the challenging problems of school accountability. However, if policymakers follow these recommendations, it will result in accountability systems that are more valid, more reliable, fairer, and more transparent. These improvements should have important effects at reducing the unintended consequences of standards-based accountability moving forward.

CEPWC Working Paper Series No. 13 July 2013.
Available at http://curry.virginia.edu/research/centers/cepwc/publications.
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

Table 1

Waiver Application Status as of May 2013

Approved		Under Review	Rejected	Not Submitted
Arizona	Missouri	Alabama	California	Montana
Arkansas	Nevada	Alaska		Nebraska
Colorado	New Jersey	Bureau of Indian Ed.		
Connecticut	New Mexico	Hawaii		
District of Columbia	New York	Illinois		
Delaware	North Carolina	Iowa		
Florida	Ohio	Maine		
Georgia	Oklahoma	New Hampshire		
Idaho	Oregon	North Dakota		
Indiana	Rhode Island	Pennsylvania		
Kansas	South Carolina	Puerto Rico		
Kentucky	South Dakota	Texas		
Louisiana	Tennessee	Vermont		
Maryland	Utah	West Virginia		
Massachusetts	Virginia	Wyoming		
Michigan	Washington			
Minnesota	Wisconsin			
Mississippi				

Note. Washington has not fully developed the index it is using to identify priority and focus schools, so it is not included in our analyses. Alabama, Alaska, Illinois, Iowa Maine, and West Virginia were allowed to modify or freeze their NCLB AMOs for 2012-13