The online version of this article can be found at:
http://edr.sagepub.com/content/43/1/45
The Waive of the Future? School Accountability in the Waiver Era

Morgan S. Polikoff¹, Andrew J. McEachin², Stephani L. Wrabel¹, and Matthew Duque¹

Forty-two states and the District of Columbia have recently received waivers to the school accountability requirements of the No Child Left Behind Act (NCLB). As the prospects for reauthorizing the Act in the near term are dim, these new accountability systems will be law for at least several years. Drawing on a four-part framework from the measurement literature, we describe and critique the approved waiver accountability plans, comparing them to the NCLB accountability rules. We find a mixed bag—some states have made large improvements and others have not. Overall we conclude that states missed opportunities to design more effective school accountability systems that might minimize negative unintended consequences of these policies. The article concludes with suggestions for state and federal policy in light of the available literature.

Keywords: accountability; policy; policy analysis; validity/reliability

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, 2007; Davidson, Reback, Rockoff, & Schwartz, 2013; Ho, 2008; Linn & Haug, 2002; Polikoff & Wrabel, 2013; Porter, Linn, & Trimble, 2005). In 2011, to help alleviate the impending 100% proficiency deadline, the federal Department of Education permitted states to apply for waivers to NCLB requirements. As of October 2013, 42 states and the District of Columbia have received waivers in exchange for creating new state accountability systems.¹

The theory of action of accountability policies posits that the use of incentives will motivate educators to align their behaviors with predetermined standards and goals (Figlio & Ladd, 2007; Smith & O’Day, 1990). However, school accountability policies rely on several criteria that affect the opportunity for these policies to elicit intended outcomes. When these criteria are violated, accountability policies often lead to unintended consequences that may harm students’ success (Figlio & Ladd, 2007; Figlio & Loebl, 2011).

In this article, we focus on the fundamental principle that accountability systems are only as good as the data on which they are based. Thus, we evaluate the ESEA waivers using four standards of practice regarding the appropriate use of assessment data established by the American Psychological Association (APA), American Education Research Association (AERA), and National Council on Measurement in Education (NCME), as well as the measurement and assessment literature: construct validity, reliability, fairness, and transparency (APA, AERA, & NCME, 1999; Baker & Linn, 2004; Kane & Staiger, 2002; Linn, 2000, 2004). These four criteria, while nonexhaustive, allow us to evaluate important conditions of accountability systems in the pre-implementation phase. Researchers have also analyzed the quality of accountability policies by examining their equity implications (Domaleski & Perie, 2013) and other dimensions. Although these are important lines of inquiry, they are not part of the APA/AERA/NCME standards and are outside the scope of this article.

In what follows, we first define these criteria and illustrate how they have played out under NCLB’s accountability system (for a brief description of accountability under NCLB, see the Appendix). Next, we evaluate the extent to which school accountability systems created in the accepted waivers violate the criteria. We hope that identifying the violations of these criteria

¹University of Southern California, Los Angeles, CA
²North Carolina State University, Raleigh, NC
early in the waiver process will give policymakers and educators a better understanding of the importance of accountability system design for affecting the likelihood of achieving desired outcomes.

**Accountability System Design**

Broadly, there are two primary streams of economic theory that support the use of accountability in education. The first is principal agent theory (Holmstrom & Costa, 1986; Milgrom, 1988; Milgrom & Roberts, 1988), which suggests that the incentives created through accountability systems help direct educators’ efforts toward those behaviors most important for improving student outcomes (Acemoglu, Kremer, & Mian, 2008; Kremer & Sarychev, 2000; Prendergast, 1999, Stein, 1988). The second is the experiential goods literature (Shapiro, 1983), which argues that the infusion of quality information helps educational consumers (e.g., parents, students) make better choices from among educational options (Charbonneau & Van Ryzin, 2011; Figlio & Loeb, 2011; Jacobsen, Sautlitz, & Snyder, 2012; Rothstein, Jacobsen, & Wilder, 2008). Although a full review of this literature is outside the scope of this article, the benefits of accountability under each theory are dependent on the specific type and quality of information provided to educators and consumers. In this section, we use the APA/AERA/NCME standards to describe four criteria that define high-quality data necessary to hold schools accountable for student outcomes and help parents and others make educational decisions.

**Construct Validity**

In our context, construct validity is the set of defensible inferences that can be established from a set of performance measurements (Crocker, 2006; Cronbach, 1955). School classifications made under an accountability policy have construct validity if the performance measures adequately cover the latent set of desired student outcomes, and if the inferences made on the basis of those measures are appropriate. Accountability policies implemented to date typically rely on objective performance measures—primarily student test scores. It is assumed that aggregate math and English language arts (ELA) scores closely proxy important unmeasured goals such as citizenship, ethics, and critical thinking (Rothstein et al., 2008).

NCLB 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 (percent proficient) does not account for schools’ contributions to student learning (Heck, 2006; Krieg & Storer, 2006; Weiss & 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 & Staiger, 2002). Third, although NCLB regulations allow states to use growth-to-proficiency models, these models do not meaningfully account for school improvement (Polikoff & Wrabel, 2013; Weiss & May, 2012). Fourth, NCLB’s focus on only mathematics and ELA proficiency falls short of capturing all the important outcomes of schools.

**Reliability**

Reliability is the consistency of a performance classification either between multiple measures at the same time (e.g., proficiency versus growth) or between the same measure at multiple time points (e.g., year-to-year stability) (Crocker & Aligna, 2006). Our work concerns the stable classification of schools across years.

Many factors can affect the reliability of performance measures; the two most important are the type/level of information and the number of years. Under NCLB, schools are primarily accountable for proficiency rates, which are highly reliable. However, because NCLB proficiency targets are rapidly increasing, many schools are in fact accountable for changes in proficiency rates, which are noisy due to measurement error and sampling variation (Kane & Staiger, 2002). The number of years used to generate the measure also affects reliability; using more years of data can improve reliability (McEachin & Polikoff, 2012).

**Fairness**

We define fairness in accountability systems as the level of disparate performance classifications of schools according to their demographics; that is, “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” (Camilli, 2006, p. 225). Another way of thinking about the fairness of a performance measure is to consider a school’s reference population. The population can be absolute (i.e., all schools) or conditional (e.g., schools with similar student populations) (Barley & Neal, 2011)—the latter would reduce the influence of non-school factors.

A fair accountability system would be one that holds schools accountable for only the portion of student achievement they can 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 (Clotfelter & Ladd, 1996). However, NCLB holds schools accountable for the percent of students who score at least proficient; thus, larger, more diverse schools (Balfanz et al., 2007; Krieg & Storer, 2006; Sims, 2013) and schools with lower initial achievement (Riddle & Kober, 2011) are more likely to fail Adequate Yearly Progress (AYP). Unless an accountability policy makes specific provisions to account for nonschool influences on achievement, the system will likely exhibit some degree of unfairness (Balfanz et al., 2007; Ehlert, Koedel, Parsons, & Podgursky, 2013).

**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, then the weights should be coherently and explicitly stated; (2) if schools are classified based on student assessments, then information about the
error rates and quality of the assessments should be public; and (3) yearly reports provided to stakeholders should promote the valid interpretation of the results from students’ assessment and school classifications.

NCLB’s accountability measures are fairly transparent; proficiency rates are more straightforward than many other achievement measures. However, the lack of a common meaning for proficiency across states (National Center for Educational Statistics, 2007) reduces transparency. Furthermore, there are numerous alternative methods to make AYP other than meeting the proficiency targets; these are not transparent, but they account for an increasingly large proportion of schools (Polikoff & Wrabel, 2013).

ESEA Flexibility

Flexibility Policy

The provisions of NCLB remain in effect because the ESEA was not reauthorized in 2007 as scheduled. However, the U.S. Department of Education (USDOE, 2011) has stated that parts of NCLB have become barriers to education reform. While waiting for reauthorization, Secretary of Education Arne Duncan offered states the opportunity to request flexibility from certain NCLB mandates in exchange for pursuing comprehensive plans to reduce achievement gaps, improve instruction, and advance outcomes. The USDOE has identified four waiver principles. The principle relevant to school accountability is “differentiated recognition, accountability and support.” The first requirement under this principle is to identify which subject areas will be assessed and used for accountability. Second, states must outline their new annual measurable objectives (AMOs), either creating their own plan or choosing between two options: AMOs that increase in equal annual increments to 100% proficiency by 2019–2020 or reduce by half the percentage of below-proficient students in the “all students” group and in each subgroup within 6 years. Third, states determine how subgroups will be included in accountability policy. Fourth, states outline how they plan to evaluate school performance using the new AMOs and performance indicators. 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 or highest progress schools are reward schools, which may not have large achievement gaps. 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 test score or graduation rate gaps. Title I high schools with graduation rates below 60% must also be identified as focus schools. Priority schools are defined as the state’s lowest performing schools; they must comprise at least 5% of the state’s Title I schools. A priority school may be identified based on the achievement or graduation rate of all students, or if it is implementing a School Improvement Grant (SIG) intervention model.

For this analysis, we examine the identification rules for priority and focus schools, because the AMOs are not necessarily used by states for accountability. Although we acknowledge that decisions about other components of the waiver applications (e.g., educator evaluation policies) may have influenced school accountability plans, we focus on school accountability because educator accountability was not a feature of NCLB. Where appropriate, we acknowledge the constraints and limitations states may have faced when making policy choices, although we do not explore why states made the choices they did. Such an investigation would be a welcome contribution to the literature.

Approved Waivers

As of October 2013, 45 states and the District of Columbia have submitted ESEA flexibility requests, as seen in Table 1. Of those requests, 43 have been approved and three—Illinois, Iowa, and Wyoming—are under review. Five states—California, Montana, Nebraska, North Dakota, and Vermont—either did not submit a waiver, were rejected, or withdrew their applications. We focus only on approved flexibility requests.

Methods

To evaluate the new accountability systems, we analyzed each approved request in a three-phase process. Two research team members reviewed each waiver, and each reviewer analyzed each waiver at least once. For the first phase, the waiver applications were read, outlined, and condensed according to the waiver principles. The second phase used the condensed outlines to code the accountability designs, including such features as sub-group size, subjects tested, components and weights of composite indices, and growth measures. If two reviewers disagreed on interpreting an aspect of a waiver application, a third reviewer was consulted. Finally, we applied the four measurement criteria to describe each waiver application. Our analysis relies only on the waiver requests as written.

Below, we summarize our results pertaining to the four dimensions. Where applicable, we provide counts of the number of states meeting certain criteria, lists of state names in tables, and examples describing particular features in greater detail. We exclude Washington state because their application did not describe the index to be used to identify priority or focus schools. Thus, there are 42 applications analyzed. Importantly, most states identify priority and focus schools in multiple ways. For instance, half of the priority schools might be based on a composite index, while the other half might be based on graduation rates. Therefore, the numbers that follow often add up to more than 42.

Results

Construct Validity

There are two main ways in which the identification of priority and focus schools in the waiver plans is, in aggregate, superior in construct validity to the way schools were identified under NCLB. The first is in the inclusion of non-test-based measures,
as seen in Table 2. For priority schools, 26 states identify high schools using graduation rates (as was true under NCLB); most often, schools with less than 60% graduation are identified. Separately, 23 states identify priority schools using a composite performance index (e.g., an A to F grade based on a combination of measures). Of these 23 indices, 19 include graduation rates and 15 include other measures not based on state tests. The most common non-state-test measure is some college/career ready indicator (12 of the 15), but states also include attendance, test participation rates, educator effectiveness, school climate, and opportunity to learn measures. Only Arkansas, New Hampshire, Pennsylvania, Wisconsin, and West Virginia use test scores alone in identifying priority schools (and Utah uses SIG schools, which are identified by low proficiency). Focus schools are more often identified using achievement test data. Still, most states use at least one non-test-based measure for identifying focus schools. All-student or subgroup graduation rates are used in 20 states. Twelve states use their composite index; 11 of these indices include graduation rates, and 9 have other non-state-test measures.

A second way in which focus and priority classifications are superior to AYP in construct validity is that many states plan to use test-based measures other than proficiency rates. Although no state uses achievement growth as its own criterion for identification, 20 states use a composite index that includes growth. The weight on the growth measure in the composite index ranges from 14% for Kentucky's high schools to 75% for Idaho's elementary schools. Some states do not assign weights to growth measures but rather count students as proficient for the purposes of calculating proficiency rates if they have passed their growth target. The vast majority of growth measures use Student Growth Percentiles (Betebenner, 2011) or some variant thereof. These measures are much closer to identifying schools' contributions to student learning than NCLB's percent proficient. Even in states where achievement levels remain an important part of the identification system, 12 are moving from proficiency rates to a system that allocates points along the achievement distribution. These systems improve over NCLB because they incentivize schools to focus on all students. Finally, for focus classifications, 21 states use either a proficiency or graduation gaps measure, directly targeting the identification and reduction of achievement gaps.

Although these are promising signs, almost anything would have been an improvement over the construct validity of AYP classifications, which were based only on proficiency rates, or changes in proficiency rates, in math and ELA. However, there are two main shortcomings with the construct validity of the proposed systems. For one, the ESEA guidelines did not require states to include science assessment in their accountability systems; thus, most states (28) are still using only mathematics and ELA to identify schools (we consider reading, language arts, and writing to all represent ELA). Of the 14 that are using other subjects for accountability, 23 states are testing these subjects in only a few grades, and most give the

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

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.
Priority or focus schools identified using performance in subjects other than reading, writing, science, and social studies. The weakest state measure proficiency using points along the distribution.

Reliability

Evaluating the reliability of priority and focus classifications is more difficult, because the systems have not been implemented. We know that AYP was highly reliable in identifying low-performing schools, because school-level proficiency rates are reliable across years (McEachin & Polikoff, 2012). On the one hand, most all of the waiver states require their focus and priority schools to be labeled as such for 2 to 3 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 waiver classifications will be less reliable than AYP. The first is that the priority and focus classifications are based on a fixed percent of schools. This norm-referenced approach results in decreased reliability due to measurement error and imprecision around the cut (Ho, 2008; Kane & Staiger, 2002). There are likely not meaningful differences in performance between a 4th percentile school and a 6th percentile school; yet, under the waivers, one would fail and the other would not. Although they might be classified as a focus school if they just miss the priority cut, this is rarely true given that most states use different measures to identify focus and priority schools.

A second reason for decreased reliability is the use of growth models in composite indices. With what is known about teacher-level measures of student growth (Goldhaber & Hansen, 2012; McCaffrey, Sass, Lockwood, & Mihaly, 2009), the year-to-year stability of school-level student growth measures is likely moderate at best. Although states could use multiple years of data for their accountability measures, only 12 states chose to do so, and some of these states used multiple years of data for their status measures only (which improves reliability marginally). Even with multiple years of data, composite indices that incorporate growth measures will be less stable than AYP.

In short, the use of growth data comes with an important tradeoff: it enhances construct validity but decreases reliability. The most reliable systems will therefore be those that rely on status measures of performance, such as those highlighted above as having weak construct validity. Among states with more construct valid measures, those using multiple years of data will have greater reliability. An example is Massachusetts, which weights...
its student growth percentiles using 4 years of data prior to inclusion in the composite index.

**Fairness**

The fairness of the waiver plans will likely be an improvement over the current AYP system; however, they will still suffer from similar biases against schools serving more students from historically low-performing subgroups because of the heavy reliance on status-based measures of achievement. Table 3 contains a set of fairness indicators. Even in states using indices to identify priority schools, (a) proficiency rates or other status measures represent 50% or more of the index for at least some grades in 16 states (and are included in the index in all states), and (b) all but three states identify low-proficiency SIG schools as priority schools. Proficiency rates also factor in the identification of focus schools: 25 states identify focus schools using either subgroup proficiency or a subgroup index based on proficiency rates, and another 12 use their composite indices, which have heavy status components. Subgroup proficiency measures target schools serving diverse populations, particularly those serving large proportions of students with disabilities (McEachin & Polikoff, 2012). Although some states have controversially (McNeil, 2012) established fairer AMOs that differ for each subgroup, these are not used for 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 measure the largest gaps between two subgroups in a school. These measures are likely unfair to diverse schools (McEachin & Polikoff, 2012), but not as unfair as status measures. The second type of proficiency gap measure compares the performance of a subgroup in a school to a state average or state-wide target. Although these are called “gap” measures, they are actually subgroup status measures.

Although diverse schools will be more likely to fail under any accountability system save one that explicitly controls for student demographics, there are some ways in which the approved waivers will decrease the diversity penalty. One way is in states using “super subgroups” rather than NCLB subgroups for priority or focus classification. Super subgroups generally take two forms. One is a combination of subgroups based on demographics (17 states). For instance, Mississippi’s super subgroup includes all students in any traditionally low-performing NCLB subgroup. The other is a subgroup of the lowest performing students in a school (nine states). For instance, Michigan’s composite index includes a gap measure that compares the achievement of the top 30% to the bottom 30% in each school. Both approaches will reduce the diversity penalty, although the latter type will reduce it more since it is agnostic to demographics. While enhancing fairness, the use of super subgroups may trade off with construct validity if the super subgroup masks low performance of any traditional subgroup. Indeed, some civil rights groups have expressed concern that the use of super subgroups in the waivers will result in students from disadvantaged subgroups not receiving the attention and support they need (e.g., Hernández, 2013).

Composite indices incorporating growth models will also be fairer than those based more heavily on status measures, because the correlations of growth measures with student characteristics are smaller. However, these correlations will not be zero, and some have expressed concerns about bias (Ehler et al., 2013). One way to control this problem would be to explicitly control for student and school demographics in a value-added model (Ehler et al., 2013; Reardon & Raudenbush, 2009), although employing such controls might be seen as setting different standards for different groups of students. However, the USDOE waiver guidelines, while allowing states to set AMOs that differ across subgroups, prohibit the use of demographics in the estimation of growth models. Thus, although most states use value-added models, SGPs, or other student-growth approaches in their composite indices, none of these models control for student demographics. These approaches, while fairer than status measures, are likely biased by student characteristics unrelated to school practices and policies (Ehler et al., 2013).

**Transparency**

Although NCLB’s AYP system was unfair and had weak construct 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 an aggregate. These indices, because of their familiar form, should be fairly interpretable by educators and the public. Many states and districts have used similar systems for years.

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
priority or focus schools, as seen in Table 4. Among the 38 states with a composite index, 12 do not use it to identify priority schools and 18 do not use it to identify focus schools. Another three states use either a modification of their index or only some portion of their index (e.g., one subscale) to identify priority schools, and 10 for focus schools. For instance, Minnesota recalculates its regular index methodology but applies it only to the historically low-performing subgroups to identify focus schools. In these cases, the schools with the lowest composite index will not necessarily be the ones identified as priority or focus, sending confusing messages to educators and the public.

Less clear still, in some states there is a third method of measuring performance for the AMOs. In Nevada 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 system may send unclear messages, because the priority and focus schools may not be the lowest performing schools on either the index or the AMOs. Only 16 states determine either priority or focus schools using the same measure as for their AMOs.

Another way in which the transparency of the measures is unclear is that many state indices apply seemingly arbitrary weights to unrelated measures to arrive at a composite score. In these cases, there is an apparent tradeoff between the increased construct validity that comes with a composite index and the transparency of the index. For instance, South Dakota’s School Performance Index is a 100-point scale. For elementary schools, 25% of the grade is based on proficiency rates, 25% on growth, 20% on attendance, 20% on educator effectiveness, and 10% on school climate. The points are allocated in proportion to the raw values—for instance, a school with 70% of educators rated “proficient” would receive 14 points for that component. Thus, although the 100-point index is conceptually transparent, it is not clear whether a school scoring, for example, 80, is effective. Furthermore, it is not immediately apparent from the index what a school might do to improve its score.

A final challenge in some indices is in the calculation of the subcomponents using contingency tables to transform continuous variables for the purposes of inclusion in the index. For instance, in Idaho each subscale in the index is first converted to 5- or 10-point subindices, 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. These kinds of approaches are conceptually unclear, and states rarely offer rationales for their use. These approaches also 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, schools could merely determine on which measure they are closest to the next point bump and target their efforts accordingly.

Discussion

As the 2013–2014 deadline of 100% proficiency looms, there has been increasing pressure on the federal government to update ESEA. Given Congress’s inaction in reauthorizing ESEA, the USDOE implemented a waiver program in 2012 to reduce the law’s burden on states. The waiver program provides states the opportunity to implement their own accountability systems, often substantially reducing the number of schools identified for intervention. Overall, the waivers provide a mixed bag of improvements over, and duplications of, the problems of NCLB.

In many of the waivers, states have strengthened the construct validity of their accountability systems by using nontest measures and measures of student growth. These changes should capture more of the multidimensional nature of schooling, increasing the alignment between incentives and desired outcomes.

In most states, however, many of NCLB’s problems are duplicated. For example, although states were allowed to test additional subjects for accountability, only 14 of the states (33%) chose to do so, and all but perhaps Michigan maintain a heavy focus on math and ELA. This decision may be motivated by concerns about cost or the amount of testing in schools. But given complaints about narrowing of the curriculum to mathematics and ELA, and given that states currently spend a fraction of 1% of education dollars on testing (Chingos, 2012), it is surprising that many states did not take the opportunity to expand on “what counts” to include other content areas. Also, the reliance on proficiency status to identify schools will continue to penalize schools serving students from historically disadvantaged groups.

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 varies...
dramatically. The measures range from NCLB-like changes in school proficiency rates to student-level Student Growth Percentiles. The USDOE’s decision to prohibit states from controlling for student demographics will likely create performance measures biased by factors unrelated to school policies and practices (Ehlerl et al., 2013). Furthermore, several states implementing growth measures are not using them to identify priority and focus schools. Finally, although states could increase the reliability of their growth measures by using multiple years of data, few made this choice.

It is likely that any accountability system will violate one or more of the four criteria. The waiver applications perfectly illustrate these tensions. Although most of the waivers proposed positive changes over the NCLB system, the net change is harder to summarize. We do not claim that all decisions we highlighted were due to a lack of information or technical capacity; rather, there are undoubtedly political considerations that shaped states’ decisions. Still, there are several politically feasible changes that would improve state systems; we close with a few such policy recommendations to mitigate the unintended consequences of the current waiver plans.

Policy Recommendations

The first and most important policy recommendation is to incorporate the lessons learned from NCLB into the implementation of the waiver applications (see Davidson et al., 2013, for a thorough review of problems associated with NCLB implementation). A number of the waiver applications propose policies that are known to pose specific problems. Moving away from the use of unadjusted proficiency rates and adding additional tested subjects to accountability would improve the construct validity of classifications considerably. Since all states are required to test science, they should include science testing results in priority and focus determinations.

To further improve the construct validity and fairness of accountability classifications, the USDOE should allow states to create more refined comparison groups for schools by conditioning on student demographics in the construction of school performance measures. States may not want to be seen as setting different targets for different groups. However, schools have had different targets for different groups for a decade through NCLB’s Safe Harbor provision (Polikoff & Wrabel, 2013), and many states opted to set USDOE-endorsed subgroup-specific AMOs under the waivers, so most states are already setting different targets for different groups. By excluding student demographics from performance measures, the system expects the same performance from all schools regardless of their student inputs, penalizing schools for factors they cannot control. Although political pressures push against the use of controls for student background, this unfairness may contribute to unintended consequences, such as teachers preferring to work in schools serving more affluent children.

Another way to improve construct validity is to move away from within-state achievement 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 within-district system would send a clear message that all students within a school deserve attention and effort (Neal & Schanzenbach, 2010).

To improve the reliability of performance classifications, states should use multiple years of data for school performance measures, especially measures incorporating student growth. By now, most states have the ability to use multiple years of data in the construction of school performance measures; there is no good reason not to.

It would also improve reliability if states moved away from the arbitrary norm-referenced approach to identifying low-performing schools encouraged by USDOE guidelines. Although setting the bar at the bottom 5% or 10% creates a more manageable sample size and likely reduces total costs associated with state interventions, 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 these schools may not meaningfully differ. Rather, schools may benefit from a clear operational definition of a low-performing school that is based on a set of performance criteria. Furthermore, extant research suggests consequential accountability systems may be more likely than non consequential systems to raise student achievement (Hanushek & Raymond, 2005). Civil rights groups have expressed concern that if only a small share of schools is threatened by the priority and focus labels, the waiver program may provide weaker incentives than NCLB (e.g., Hernández, 2013).

To improve both transparency and construct validity of classification systems, states should reevaluate the construction of composite measures and their use for identifying schools. Although A-to-F systems are, on the surface, transparent, the underlying design of these systems involves a great deal of arbitrariness that makes it difficult for educators and parents to understand performance. Furthermore, keeping indicators separate allows for a more nuanced understanding of the strengths and weaknesses of schools that can be used to tailor interventions.

Finally, states should conduct short-term analyses of the implementation of their waiver systems and make adjustments. 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.

Appendix

NCLB

In the most recent federal implementation of educational accountability, Congress reauthorized ESEA with the NCLB Act of 2001, which requires states to implement a set of school accountability mandates in order to receive federal Title I 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. AMOs are the annual interim proficiency targets schools must meet in both mathematics and ELA en route to the 2014 target.
Schools are accountable for performance disaggregated by numerically significant subgroups based on racial/ethnic, disability, socioeconomic, and English language proficiency status. The federal government gave 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; Davidson et al., 2013; Porter et al., 2005).

In addition to meeting minimum proficiency targets, schools are also held accountable for a 95% 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. Although 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 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.

NOTES

1We do not include the waiver received by the California Office to Reform Education (CORE) districts, because this article is focused on state accountability.

2All information taken from USDOE, ESEA Flexibility (http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html).

3Student growth percentile is a percentile rank of students’ achievement growth conditional on at least 1 year of prior achievement. To use SGPs in accountability, states generally take the median SGP of a school’s or teacher’s students.

REFERENCES


McEachin, A., & Polikoff, M. S. (2012). We are the 5%: Which schools would be held accountable under a proposed revision of the Elementary and Secondary Education Act? Educational Researcher, 41(7), 243–251.


Polikoff, M. S., & Wrabel, S. L. (2013). When is 100% not 100%? The use of safe harbor to make Adequate Yearly Progress. Education Finance and Policy, 8(2), 251–270.


AUTHORS

MORGAN S. POLIKOFF, PhD, is an assistant professor at the University of Southern California Rossier School of Education, 3470 Trousdale Parkway, WPH 904A, Los Angeles, CA 90089; polikoff@usc.edu. His research focuses on the design and effects of standards, assessment, and accountability policies.

ANDREW J. MCEACHIN, PhD, is an assistant professor of Education Policy and Program Evaluation at North Carolina State University, College of Education, 608E Poe Hall, Campus Box 7801, Raleigh, NC 27695-7801; ajmceach@ncsu.edu. His research focuses on the design and impact of accountability systems, math education policy, and the economics of education.

STEPHANI L. WRABEL is a dean’s fellow at the University of Southern California, Rossier School of Education, 3470 Trousdale Parkway, 901D Waite Phillips Hall, Los Angeles, CA 90089; swrabel@usc.edu. Her research focuses on accountability policy, nonpromotional school transitions, and military-connected students and schools.

MATTHEW DUQUE is a PhD candidate at the University of Southern California, Rossier School of Education, 3470 Trousdale Parkway, 500 Waite Phillips Hall, Los Angeles, CA 90089. His research focuses on the distribution of teacher quality within and across schools.

Manuscript received May 8, 2013
Revision received October 22, 2013
Accepted October 29, 2013