Article

Educational Accountability and State ESSA Plans

Educational Policy
2022, Vol. 36(3) 717–747

DOI: 10.1177/0895904820917364
journals.sagepub.com/home/epx
SAGE

John Portz[1] and Nicholas Beauchamp[1]

Abstract

This paper examines different state approaches to educational accountability in response to the Every Student Succeeds Act. Cluster analysis reveals three groups of states with similar indicator weights and rating systems, and principal component analysis identifies two dimensions underlying these clusters. We find that state-level demographics are correlated with the types of assessment policies adopted by states: policy liberalism is associated with putting greater weight on school quality and student success, while economic variables are associated with traditional performance measures, such as graduation rates and testing. These clusters reveal different approaches to measuring accountability and prioritizing different kinds of information, which can in turn influence the nature of education politics.

Keywords

accountability, education policy, politics of education, state policies, cluster analysis, principal component analysis

Introduction

In December 2015, President Obama signed the Every Student Succeeds Act (ESSA), ending the 13-year run for the No Child Left Behind Act (NCLB). With this long-awaited legislative change, the political pendulum on K-12 school accountability swung back toward the states. While there

[1]Northeastern University, Boston, MA, USA

Corresponding Author:
John Portz, Northeastern University, 360 Huntington Ave., 922 Renaissance Park, Boston, MA 02115, USA
Email: j.portz@northeastern.edu

are numerous examples of continuity and change between the two federal acts, one significant and often-noted change is more flexibility, albeit within certain parameters, for state policymakers and educators to design various parts of their K-12 education policies, including their accountability systems (Hess & Eden, 2017). As one scholar notes, "ESSA provides states and localities with substantial new flexibility to pursue reform grounded in local priorities" (Jochim, 2017, p. 135).

Thus began a period of state adaptation, innovation, and redesign. In this context, our focus is on two key areas of state accountability systems: accountability indicators and the overall rating assigned to schools. Indicators, such as student achievement and high school graduation, are at the center of an accountability system. They are the means by which policymakers, the public, and other "account-holders" determine whether students, schools, and other "account-givers" are meeting identified goals. We also focus on the different types of overall accountability ratings that states assign to schools. These ratings, such as an A–F grade or "exemplary" performance level, summarize the performance of a school and are important markers for policymakers, educators, and the general public. In both areas—indicators and overall ratings—states have several key points of flexibility to craft an accountability design that fits their particular interests and preferences.

Scholars and others are split on how creative states will be in this process. Some are skeptical. Arnold Shober (2017), for example, notes how "uncreative" states were in designing indicators under the flexibility of NCLB waivers granted by the Obama administration (p. 109; see also McNeil, 2014). In a similar vein, Cynthia Brown (2017) highlights the limited capacity of most states to truly engage in robust policymaking and, importantly, policy design and implementation. Others are more optimistic. Martin West (2017), for example, contends that ESSA strikes a good balance by giving states "both space and incentive for experimentation," such as incorporating both growth and status in their use of student test scores (p. 82). From within government itself, a majority (25 or 34) of state education agencies responding to a survey indicated they had "sufficient capacity to measure and report on ESSA's accountability indicators" (Center on Education Policy, 2017, p. 10).

Our analysis of accountability indicators and overall ratings contributes to this debate by focusing on three questions:

- *What are distinctive state responses in the design of accountability indicators?* There are both commonalities and differences in state designs. We explore this terrain by identifying three clusters of state responses along with two major underlying dimensions using cluster analysis and principal component analysis (PCA).

- *What factors or variables help to explain these different responses?* We consider economic, political, motivational, and institutional variables to help explain the different response patterns. Our analysis points particularly to the importance of economic variables regarding spending and poverty and a political variable represented by policy liberalism.
- *What are the consequences for education politics?* The information provided in an accountability system has important interpretive and resource effects that shape education politics. The three clusters identified here offer different kinds of information that shape the political dialogue in favor of some interests over others.

Research Design

This research is based on a comparative analysis of state plans submitted under ESSA along with several statistical tools to analyze similarities and differences in the accountability aspects of those plans. All 50 states and the District of Columbia submitted plans to the federal government, as required by ESSA, that describe how each intends to implement various programs and services under the new federal law. Included in each plan is a section on accountability, which covers the following key topics: long-term goals, student subgroups, indicators, classification of schools, and intervention strategies. The federal Department of Education provided a template for states to follow in preparing their plan (See Supplemental Appendix A). In March and November of 2017, states submitted plans, and all were approved by July 2018.

The analysis in this article includes a review of 51 plans (includes Washington, D.C.) as well as material from several organizations that reviewed the plans, including the Education Commission of the States, Achieve, and Education Evolving. Phone calls and e-mail correspondence with nonprofit policy experts and state officials also contributed to the analysis. In our analysis of accountability indicators and overall ratings, we employ two fundamental approaches from unsupervised machine learning: cluster analysis and PCA (James et al., 2013). We use these approaches to identify three clusters of states with common policy designs and to reveal the underlying dimensions that structure these clusters.

We then turn to a set of independent variables and regression analysis to help explain the groupings and dimensions derived from cluster analysis and PCA. The choice of variables draws, in part, from the state innovation and diffusion literature. Frances Berry and William Berry (2018), for example, in their focus on internal determinants of innovation, highlight motivation, resources, learning, as well as political and institutional variables. Adapting

these categories to our study, we focus on four general variables—economic, political, motivational, and institutional—that can influence how state policymakers assign weights to accountability indicators and develop an overall accountability rating.

An *economic* variable focuses on the resources available to policymakers. A strong state financial position, in particular, can facilitate a more robust process of policy debate and innovation (Boehmke & Skinner, 2012). Choosing variables relevant in an educational setting, we use two measures to assess state economic status: the percent of students in poverty and per pupil expenditures. A *political* variable provides a different lens that focuses on the ideological orientation of state policymakers to the development and innovation of public policies. In this regard, more liberal states tend to be more innovative (Boehmke & Skinner, 2012; Nice, 1994). Two measures represent this variable: policy liberalism and political party of the governor.

Motivation is a third variable that influences how policymakers shape policy designs. Are policymakers motivated to add to or alter the design of their accountability indicators? Motivation is derived from a number of sources, including the presence and complexity of a perceived problem (Baumgartner & Jones, 2015; Boushey, 2010). In our analysis, we consider student test scores on the National Assessment of Educational Progress (NAEP) as a measure that might motivate educational policymakers. We also include an *institutional* variable that captures the persistence of existing policy practices to continue into the future (Pierson, 2004). For this measure, we compare the use of accountability indicators by states *prior* to passage of ESSA and then in their ESSA plans. This variable is not included in the statistical analysis, but provides additional support for identified trends (see Supplemental Appendices B and C for more detail on the independent variables).

From NCLB to Waivers to ESSA

Importantly, ESSA was not written on a blank slate. Rather, its immediate predecessor, NCLB, signed by President Bush in 2002, reinforced a growing role for the federal government in the design and implementation of accountability systems (Manna, 2011; McGuinn, 2006). Among its key accountability provisions, NCLB set a goal of proficiency for all students in English language arts and mathematics by 2014. This included all statutorily defined student subgroups, such as racial and ethnic minorities and students with disabilities. States, however, retained responsibility for developing curriculum standards, preparing test instruments, and defining proficiency.

Academic achievement on student tests, aggregated to the school level, was the primary indicator to monitor progress in meeting the proficiency goal.[1] This was a test-based system in which schools tracked student test

scores to meet "adequate yearly progress" (AYP) necessary to reach the 2014 goal. Schools that did not meet AYP were deemed "in need of improvement" and subject to escalating consequences for each year they remained off-track.

By 2011, it was clear that many schools would not meet the 2014 proficiency goal. Citing the failure of Congress to revise the law, President Obama offered states a waiver process to remove key NCLB accountability provisions, such as AYP and the 2014 proficiency requirement, in exchange for adopting education policies advocated by the administration, including college and career-ready expectations for all students; differentiated recognition and accountability; and support for effective instruction and leadership (U.S. DOE, 2013). States quickly lined-up to apply. After several application rounds, waivers were approved for 42 states, the District of Columbia, and a consortium of eight school districts in California (Center on Education Policy, 2012a, 2012b; Martin et al., 2016).

During this waiver-driven process of redesign, several developments are of particular note in shaping subsequent strategies under ESSA. One was the increasing popularity of growth models to analyze student test scores (Data Quality Campaign, 2019; Martin et al., 2016). Historically, NCLB focused on achievement (or status) test scores, rather than growth, to measure a student's performance at a *single point* in time. Increasingly, however, many states and school districts also turned to growth models that focused on changes in student test scores *over time.* A growth perspective gained popularity as a way to focus on student improvement rather than continually describing students and schools as failing to reach proficiency-based benchmarks.

A second area of growing interest was the use of non-test indicators, such as attendance and measures of school climate, to capture a broader perspective on teaching and learning (Council of Chief State School Officers, 2011; Darling-Hammond et al., 2014, 2016; Darling-Hammond & Hill, 2015; Hamilton et al., 2013; Schneider, 2017). At a limited level, this had been a part of NCLB from the start, as evident by the attention to graduation rates, but it was overshadowed by the dominant focus on test scores. The waiver process gave clear attention to college and career readiness as an indicator, and states began to explore other measures and metrics to assess student and school performance.

NCLB waivers marked a shift in many state accountability systems, but they also added to the tension on Capitol Hill. The waiver process, driven by the executive branch, slowly eroded support from a wide range of constituencies, including Republicans and some other members of Congress who saw the waiver process as an "end-run" by the Obama Administration around the law (Saultz et al., 2017). By 2015, with Republicans controlling Congress, any replacement of NCLB needed to meet Republican demands to reduce the federal government's role and provide more state flexibility in crafting education policy, including in

the area of accountability. Seeking a path forward, the Obama administration agreed to a variety of provisions that met Republican demands. Passage of ESSA in 2015 was the outcome of this compromise (Klein, 2017).

Indicators: Expanding the Terrain

Under the new law, each state prepares an "ESSA plan" describing how it will meet various provisions of the act. In the accountability part of this plan, states identify goals and indicators to monitor and assess school performance in the following areas:

- Academic *achievement* in all public schools as measured by proficiency on annual assessments, with the option of including student growth at the high school level.
- Student *growth* on annual assessments at the elementary and middle school level or another academic indicator that allows meaningful differentiation in school performance.
- High school 4-year *graduation rate*, with the option of extended-year rates.
- *English language proficiency.*
- *School quality or student success* (SQ/SS) as measured by a state-chosen indicator.

In addition to describing each indicator, the state plan includes the relative weight given to each at the elementary and secondary levels. Indicator weights at the elementary level total to 100%, as do indicator weights at the high school level. Indicator weights, however, are subject to an important limitation. Weights for assessing academic achievement, academic growth (if used), graduation rate, and English language proficiency (first four bullets listed above) must each be given "substantial weight" and, in the aggregate, these indicators must have "much greater weight" than the indicator for SQ/SS (fifth bullet). With these limitations, ESSA ensures an emphasis on test scores, proficiency assessments, and graduation rates as the long-standing, traditional indicators.

Indicator weights are listed in Table 1 for each state at the elementary and high school levels. Importantly, in a number of cases, these weights are adjusted by the authors (noted in brackets) from original state submissions to account for state-specific practices and to assign activities in a way that reserves the SQ/SS indicator for non-test measures and metrics (see footnote to Table 1). As a result of these reallocations and adjustments, the ESSA restrictions on "substantial weight" for each of the first four indicators and

Table 1. Indicator Weights (%) and Overall Rating by State.

State	EL achievement	EL growth/ other indicator	EL English learners	EL SQ/ SS	HS achievement	HS growth/ other indicator	HS English learners	HS graduation	HS SQ/ SS	Overall rating
AL	40	40	5	15	20	25	5	30	20	A–F
AK	(35) 30	40	15	(10) 15	60	0	10	20	10	Performance Levels
AZ	30	50	10	10	30	20	10	20	20	A–F
AR	(40) 35	42	8	(10) 15	(38) 35	29	6	15	(12) 15	A–F
CA[a]	20	20	20	40	17	17	17	17	33	Indicator Dashboard
CO	(32) 23	(48) 40	(12) 20	(8) 17	(30) 20	32	8	15	(15) 25	Performance Levels
CT	32	42	11	16	52	0	7	13	29	Performance Levels
DE	(43) 30	40	10	(7) 20	(44) 40	(10) 0	10	15	(21) 35	Performance Levels
DC	30	40	5	25	40	0	5	(20) 11	(35) 44	1–5 Stars
FL	(38) 25	50	12	(0) 12	(36) 18	36	9	(9) 7	(9) 29	A–F
GA	(37) 30	(39) 32	4	(20) 35	(62) 47	(5) 0	3	15	(15) 35	Performance Levels
HI	40	40	10	10	30	0	10	50	10	Support Levels
ID	36	36	18	10	45	0	23	23	10	Indicator Dashboard
IL	20	50	5	25	20	0	5	50	25	Performance Levels
IN	43	43	10	5	15	15	10	30	30	A–F
IA[b]	26	44	10	10	52	0	10	15	13	Support Levels
KS	(50) 25	25	25	(0) 25	(40) 20	20	20	20	(0) 20	Performance Levels
KY[c]	40	27	13	20	40	15	10	10	25	1–5 Stars
LA[d]	(64) 50	25	(11) 0	(0) 25	21	0	(4) 0	42	(33) 37	A–F
ME	42	38	10	10	40	0	10	40	10	Performance Levels
MD	20	35	10	35	30	(0) 10	10	15	(45) 35	1–5 Stars
MA	60	20	10	10	40	20	10	20	10	Performance Levels
MI	32	38	11	19	29	34	10	10	17	Support Levels

(continued)

Table 1. (continued)

State	EL achievement	EL growth/ other indicator	EL English learners	EL SQ/ SS	HS achievement	HS growth/ other indicator	HS English learners	HS graduation	HS SQ/ SS	Overall rating
MN[e]	40	30	20	10	40	0	20	30	10	Support Levels
MS	(41) 28	54	5	(0) 14	(30) 20	38	5	19	(9) 19	A–F
MO	40	30	20	10	40	0	20	30	10	Support Levels
MT	(30) 25	(35) 30	10	(25) 35	30	0	10	25	35	Performance Levels
NE[e]	40	30	20	10	20	20	20	30	10	Performance Levels
NV	(25) 20	55	10	(10) 15	(25) 20	0	10	30	(35) 40	1–5 Stars
NH[e]	35	(45) 35	20	(0) 10	35	0	20	35	10	Support Levels
NJ	30	40	20	10	30	0	20	40	10	Performance Levels
NM	(38) 30	(42) 40	10	(10) 20	(30) 25	30	5	(13) 10	(22) 30	A–F
NY[e]	35	35	20	10	35	0	20	30	15	Performance Levels
NC[f]	60	(20) 0	20	(0) 20	40	20	10	10	20	A–F
ND	30	30	10	30	25	0	10	24	41	Indicator Dashboard
OH	(34) 27	(53) 21	11	(2) 41	(22) 40	(37) 0	5	18	(18) 38	A–F
OK	39	33	17	11	50	0	17	11	22	A–F
OR	22	44	22	11	22	0	22	(33) 22	(22) 33	Indicator Dashboard
PA[e]	35	35	20	10	30	30	15	15	10	Indicator Dashboard
RI[f]	(42) 24	18	12	(27) 45	(37) 21	16	11	13	(24) 39	1–5 Stars
SC	(45) 35	35	10	(10) 20	(35) 25	0	10	25	(30) 40	Performance Levels
SD	40	40	10	10	40	(0) 25	10	(26) 13	(25) 13	Performance Levels
TN	(38) 45	(42) 35	10	10	(26) 30	(29) 25	10	5	30	A–F
TX	(50) 30	(40) 50	10	(0) 10	50	0	10	10	30	A–F
UT	(37) 25	(54) 42	9	(0) 24	(25) 17	(36) 28	6	11	(22) 38	A–F
VT	40	40	10	10	23	22	10	20	25	Indicator Dashboard

(continued)

Table 1. (continued)

State	EL achievement	EL growth/ other indicator	EL English learners	EL SQ/ SS	HS achievement	HS growth/ other indicator	HS English learners	HS graduation	HS SQ/ SS	Overall rating
VA[e]	35	35	20	10	25	25	20	20	10	Performance Levels
WA	40	50	5	5	30	0	5	50	15	Support Levels
WV	28	28	14	29	25	0	13	25	38	Indicator Dashboard
WI	38	38	10	15	38	0	10	38	15	Performance Levels
WY	25	(50) 25	25	(0) 25	20	20	20	20	20	Performance Levels
Average	*(37) 33*	*(38) 36*	*(13) 13*	*(12) 18*	*(34) 32*	*(12) 11*	*(11) 11*	*(23) 22*	*(20) 24*	

Note. Data reported in Table 1 include indicator weights submitted by states as well as weights adjusted by the authors, which appear in brackets. Adjustments are made for a number of reasons, some of which are noted below the table. First, seeking to separate test-related from non-test-related measures, the authors include all standardized test scores in the Achievement and Growth indicator categories. State ESSA plans generally do this with respect to math and ELA tests, but at the prompting of the federal Department of Education, many states place social studies and science test scores at a relatively low percentage in the SQ/SS indicator. These weights are re-allocated by the authors to Achievement and Growth. Second, several state plans are not explicit on the distinction between achievement and growth measures for test scores. Allocation between these two categories is made based on how the state describes the measure. Third, several state plans include extended graduation metrics (beyond 4 years of high school) in SQ/SS, while others put these in the Graduation Rate indicator. In Table 1, these are all in the Graduation category. And fourth, several state plans use a progressive decision rule or stages approach to apply the indicators to a school's performance. In this approach, the weight given to each indicator is variable depending upon how the school performs at each stage or decision point. In these cases, as state plans do not provide set indicator weights, we approximate an average weight based on the priority given to the indicator in the decision rule or stage process. EL = elementary; SQ/SS = school quality or student success; HS = high school; ESSA = Every Student Succeeds Act; ELA = English Language Arts.

[a]Equal weight to status and change for each of six measures; weights estimated by authors.

[b]Participation in assessments counts as 10% at each level.

[c]Each indicator has a range for its weights.

[d]State includes English learner in achievement index.

[e]"Stages" or "decision rules" used by state; weights estimated by authors.

[f]Partial point system; weights estimated by authors.

Source. Education Commission of the States (2018) and Authors.

"much greater weight" for those four compared to the SQ/SS indicator are violated in several cases. The last column in Table 1 includes the Overall Rating system, which is described below.

Achievement

Achievement columns (elementary and high school) record the relative weight given to student achievement (status) test scores in the overall rating for a school. As noted in Table 1, the average weight for adjusted measures is 37% at the elementary level and 34% at the high school level. However, there is wide variation across the states. At the elementary level, for example, the variation is from a low of 20% in California, Illinois, and Maryland to 64% in Louisiana.

Growth (or Other Indicator)

Growth columns capture the weight given to test scores that assess a student's improvement (or decline) over time. Although ESSA does not require inclusion of a growth perspective, growth scores are often used to acknowledge improvement rather than only a proficiency bar. Indeed, under NCLB waivers or in their own state accountability systems, 46 states used a growth metric for measuring accountability, particularly at the elementary level (Martin et al., 2016). At an average weight of 38%, growth scores are popular at the elementary level, although less popular at the high school level, with an average weight of 12%. As with the Achievement indicator, there is wide variation in relative weight among the states. This is particularly true at the elementary level, with a low of 18% in Rhode Island and high of 55% in Nevada. Whether a state emphasizes growth or achievement is a key decision each must make and is reflected in the cluster analysis presented below (D'Brot, 2017).

English Language Learners

Under NCLB, English language learners were incorporated into the accountability system as a subgroup. Under ESSA, this group has its own standing as a required indicator. In their plans, states identify curriculum approaches to help English learners achieve language proficiency, and they propose how to track and assess the performance of this group of students. The weighting for this indicator varies, ranging from a low of 3% for high school English learners in Georgia to a high of 25% for elementary English learners in Wyoming. The average is 13% at the elementary level and 11% at the high school level.

Graduation Rates

High school graduation is a well-known and often used indicator in accountability systems. ESSA requires states to establish graduation goals for all students and subgroups and use the "four-year adjusted cohort" calculation as the measure to report on success in meeting goals. In addition, a number of states provide measures for extended graduation rates covering students who require more than 4 years to complete high school. The average weight for this indicator is 23%, with a low of 5% in Tennessee to a high of 50% in Hawaii, Illinois, and Washington.

SQ/SS

This is ESSA's nod to the growing interest in developing non-test indicators to assess teaching and learning. The law provides several examples of indicators, such as student engagement, postsecondary readiness, and school climate. States can use other indicators, provided they are "valid, reliable, comparable, and statewide" and allow for "meaningful differentiation in school performance" across the state. Importantly, as noted earlier, greater weight, in the aggregate, must be given to the other four indicators listed in the law in comparison to this new indicator.

This indicator moves accountability systems closer to a "multiple measures" approach. However, there is wide variability across the states, which is apparent in two ways: by the weight assigned to SQ/SS measures and by the number and type of measures included. With respect to weight, the average given to this indicator is 20% at the high school level and 12% at the elementary level (see Table 1). The range, however, is quite broad. Several states give no weight to this indicator (based on our calculations), while at the elementary level California allocates 40% weighting (again, based on our calculations), and at the high school level Maryland gives this indicator 45% weighting.

With respect to number and type, states have identified from one to five measures or metrics to fit under the SQ/SS indicator. The most popular measures are chronic absenteeism, used in 37 states, and college and career readiness, used in 34 states (Education Commission of the States, 2018; Kaput, 2018; Kostyo et al., 2018; Rafa, 2017). Chronic absenteeism is typically defined by a student missing 10% or more of a school year, and college and career readiness is a broader measure that includes such metrics as academic courses completed, ACT/SAT participation and scores, advanced placement or international baccalaureate participation and scores, and career training and certificates (Achieve, 2018; Klein, 2019). Both measures have a track

record with the states. Under pre-ESSA accountability systems, chronic absenteeism was used in five states and variations of reporting attendance in an additional 18 states (Martin et al., 2016). College and career readiness was even more prominent in this pre-ESSA period; it was part of accountability systems in 30 states (Martin et al., 2016).

Overall Rating

A school's Overall Rating is the third key area in our analysis of indicators, along with the Elementary weights and High School weights. ESSA requires states to establish a system for "annual meaningful differentiation" based on the indicators in its accountability system. Specific labels or ranking systems are not specified; however, the state must identify schools for improvement and support if they fall into one of several categories, such as the lowest performing 5% of schools receiving federal funds (Lyons et al., 2017). This leaves the overall design of the rating system to the states, as long as it meets the "meaningful differentiation" standard and identifies schools for improvement and support.

An overall rating system is not new. Whether under NCLB or a separate state accountability system, states typically assigned an overall rating to schools. In doing so, a key point of distinction was use of a single label, such as a letter grade, or multiple measures to summarize a school's performance. Following this distinction, we divide the states into two broad groups: 36 states and the District of Columbia use a single, summary label, while 14 states follow a multiple measures approach.

States in the first group—36 states and District of Columbia using a summative label—employ a variety of methods to aggregate across the indicator categories described above. Typically, an index or point system is used in conjunction with the weights assigned to each indicator. Our analysis identifies three subsets within this group that vary by the nature of the rating. One subset includes 14 states that use an A–F letter grade system. For these states, a letter grade offers a clear, simple, and familiar metric. As described in the Utah plan, this approach provides "meaningful [and] readily interpretable differentiation among schools" (Utah ESSA Plan, 2017, p. 33). A second, smaller subset includes four states and the District of Columbia that use 1–5 Stars to label individual schools based on performance. And finally, a third subset includes 18 states using a state-determined Performance Level label.[2] The label, such as "exemplary," 'underperforming,' or another designation, could be based on points earned, a percentile ranking among schools, meeting targets, or another criterion. The distinction with this group is the use of a unique, state-designated summary label based on criterion relevant to each state.

States in the second general group—14 states not using a summative label—have common ground in relying upon multiple measures as opposed

to a summary label. These states fall into two subsets. The first consists of seven states we identify with the title Support Levels. In these states, the designation used is the one required by the federal government under ESSA, which calls for comprehensive or targeted support, depending upon how the school meets ESSA's criteria of performance. As all states are required to make these support designations, states in this group use a minimal approach to classifying their schools. The other seven states in this group use an Indicator Dashboard and explicitly eschew a summary label. They focus on providing multiple data points rather than aggregating data into a single rating. In the Idaho plan, for example, it is noted that

> Idaho's stakeholders were outspoken in their opposition to a summative rating for each school. It was felt that the complex calculations required to produce a summative score are not transparent, sometimes misleading, and result in a system that is not useful for parents and educators. (Idaho ESSA Plan, 2018, p. 25)

The Overall Rating is an important design feature with high visibility among policymakers, parents, and the broader community. It can be controversial. Schools labeled with a single metric, a grade of "C," for example, may face more intense scrutiny than those portrayed on a multi-indicator dashboard with shortcomings in some areas and strengths in others. A single label can send a powerful message. Possible controversy around overall ratings is a point further developed in the next section.

These indicators and overall ratings play a key role in state accountability systems. They provide an important window on what states value in public education. In this regard, we posit that states tend toward several key combinations of indicators, such as a relative emphasis on the new SQ/SS indicator or a preference for the traditional indicators of Achievement and Growth, along with a tendency to align with states using similar Overall Ratings. Furthermore, in identifying these common designs below, we look to underlying state-level variables, particularly of an economic and political nature, that can help to explain groupings of states.

Analyzing Accountability Designs

This section of this article returns to our three research questions:

- What are distinctive state responses in the design of accountability indicators?
- What factors or variables help to explain these different responses?
- What are the consequences for education politics?

What Are Distinctive State Responses in the Design of Accountability Indicators?

The first question raises the possibility of common patterns among the states in their design of accountability indicators. In essence, the focus turns to identifying common groupings of states across the three variables: indicator weights at the high school and elementary levels as well as the overall ratings. Using cluster analysis, we identify three clusters of states, each revealing a different area of emphasis with respect to the indicator weights and overall rating. To understand the policy and state characteristics that might give rise to this clustering, we employ PCA (a form of factor analysis) to detect two underlying dimensions that reveal how clusters differ and how states vary within clusters. Cluster analysis and factor analysis are the two foundational approaches of unsupervised machine learning: the first infers underlying categorical features, while the second infers underlying continuous features. In this case, both approaches are complementary: the first detects basic groups of related states and the second helps explain what policy choices and state conditions give rise to those clusters. Each of these approaches encompasses a broad family of different techniques. We employ the most well-established methods for each: K-means analysis for clustering and PCA for detecting underlying factors (James et al., 2013). Together, the results from these two methods allow us to see both which clusters states fall into and what the underlying dimensions are that affect the relationships between clusters and among states within clusters.

For both k-means and principal component analyses, we begin by creating a matrix of similarities between states, where elementary weights, high school weights, and overall ratings contribute equally to calculating the overall similarity of each state's accountability system to each other. Given this similarity matrix and a specific number of clusters, the K-means algorithm iteratively assigns clusters to states and states to clusters until it finds the optimal assignment such that states within a given cluster are most similar to each other and most dissimilar to states in other clusters. To determine the best number of clusters, we test all cluster sizes between 2 and 10 and use the "silhouette" metric to evaluate clusters by how internally similar they are relative to their similarity across clusters (see Supplemental Appendix D for details).

This process suggests three clusters as optimal, as listed in Table 2. Clusters identify important common elements for a group of states, but it is important to recognize a certain level of contingency to a state's cluster assignment. A number of states fall somewhat between clusters, and of course accountability systems can change over time, leading to shifts in cluster

Table 2. Average Indicator Weights (%) by Cluster.

Cluster	States	EL achievement	EL growth	EL English learners	EL SQ/SS	HS achievement	HS growth	HS English learners	HS graduation	HS SQ/SS	Overall rating
Cluster 1	AL, AR, AZ, FL, IN, MI, MS, NC, NM, OH, OK, PA, TN, TX, UT, VT	40	42	11	8	31	25	9	15	20	A–F Grade—13 1–5 Stars—0 Performance—0 Support Levels—1 Dashboard—2
Cluster 2	CA, DC, IL, KY, LA, MD, ND, NV, OR, RI, WV	31	34	12	23	27	4	11	25	32	A–F Grade—1 1–5 Stars—5 Performance—1 Support Levels—0 Dashboard—4
Cluster 3	AK, CO, CT, DE, GA, HI, ID, IA, KS, MA, ME, MN, MO, MT, NE, NH, NJ, NY, SC, *SD*, VA, WA, WI, WY	38	38	14	10	38	6	14	27	15	A–F Grade—0 1–5 Stars—0 Performance—17 Support Levels—6 Dashboard—1
Average		37	38	13	12	34	12	11	23	20	

Note. EL = elementary; HS = high school.

assignments and the relationships among accountability systems and other variables. So, while conceptually tidy, the cluster assignments should be seen as more varied and contingent than the hard assignments to categories might suggest, an aspect we explore further in our subsequent PCA.

Examining the cluster assignments in Table 2, a number of substantive patterns are easily discerned. Cluster 1 states focus particularly on Growth test scores as well as Elementary Achievement, while predominately employing the A–F Overall Rating system (13 of 16 states in this cluster). The focus on Growth is most evident at the high school level with 25% allocated, compared to only 4% and 6% in the other clusters. Mississippi, Utah, and Ohio are Cluster 1 states that allocate the greatest weight to student growth scores. States in this cluster also allocate slightly higher weight than the average to Elementary Achievement, but are below average on almost all of the other indicators. Test scores are clearly an important part of accountability systems for these states. Indeed, at the elementary level, test scores, Achievement and Growth, constitute 82% of indicator weights among the states in this cluster.

Cluster 2 is distinctive in its focus on the SQ/SS indicator. At the elementary level, 23% of indicator weight is given to SQ/SS, compared to 8% and 10% in the other clusters, and at the high school level, 32% is allocated to SQ/SS, compared to 20% and 15% in the other clusters. In addition, it has the lowest cluster weighting with respect to Elementary and High School Achievement. Maryland, California, and North Dakota are the leaders in this category. Cluster 2 includes states from four of the five Overall Rating categories, but most represented are those using the 1–5 Star system (5 of the 11 states in this cluster) and Indicator Dashboard (4 of the 11 states).

Finally, Cluster 3 is an eclectic group of states that give greater weight to English Learners and Graduation, along with High School Achievement. Kansas, Minnesota, and Missouri are the leaders in allocating the greatest weight across these indicators. Of these 24 states, 17 fit into the Performance Level rating category and most of the Support Level states are in this cluster as well. These are states that design their Overall Rating to fit particular state criteria, in the case of Performance Level, or choose to minimize rating labels and use only the federal intervention labels, in the case of Support Level. In general, these states follow a middle path in their weighting of indicators by continuing a long-standing and traditional focus on Achievement and Graduation, but allocating a higher than average weight to English Learners.

As mentioned above, underlying the coarse assignment of states to clusters is a more nuanced space of variation that affects how the clusters interrelate and how states are positioned within or between clusters. We use PCA of the accountability data to discern the underlying continuous factors that explain much of the variation in weights and rating types. Furthermore, as we

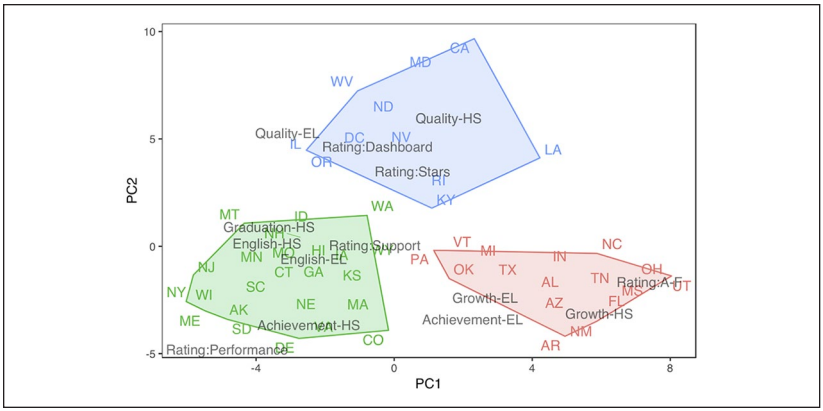

Figure 1. States, indicators, and ratings by cluster and principal component analysis.
Note. EL = elementary; HS = high school.

describe in the following, these underlying factors correlate significantly with political, economic, and other state features that may shape accountability design decisions.

PCA begins with the same state similarity matrix used for the cluster analysis, but in this case underlying continuous dimensions are found by calculating the leading eigenvectors of that matrix. As with cluster analysis, one can select any number of factors to preserve, but scree plots (see Supplemental Appendix D) suggest that, consistent with our three-cluster typology, the PCA analysis reveals two main underlying components that account for most of the variance in our data (two dimensions are the minimum required to array three clusters in a space). Figure 1 shows both k-means cluster assignments of states (color-coded in Figure 1) and each state's position in a two-dimensional space representing the first and second principal component scores assigned by PCA to each state (note that the units are arbitrary, and only relative placement matters). Because PCA analysis scores variables (policy weights) as well as units (states) on a similar scale, it also allows us to place the accountability variables—each weight and the overall rating type—in the same space, generally placing each variable closest to the states that most emphasize that weight or rating feature. While it may appear that nearby states are simply clustered together, the k-means and PCA approaches are fundamentally different, and the fact that similarly colored (clustered) states appear in close proximity to each other shows that these two different approaches are inferring very similar underlying structures.

As can be seen in Figure 1, PCA scores place Cluster 1 states in the lower right quadrant as well as the indicators of Growth and Elementary Achievement, along with the A–F rating system. This spatial placement is consistent with the cluster analysis. As principal components are continuous variables with opposed positive and negative directions (though the sign of the direction is arbitrary, as is the numeric scale), the positioning of Cluster 3 to the left of Cluster 1 indicates that the variables most highly weighted in Cluster 3 are strongly negatively correlated with those variables weighted most in Cluster 1. Cluster 3 states focus more on English Learners, Graduation, and High School Achievement, along with a preference for Performance Level ratings. The opposition of these weights to those in Cluster 1 goes beyond the mere fact that weightings are inherently in a trade-off due to necessarily summing to 100% at each level (elementary and high school). Rather, the positions of these weights and states on opposite ends of the first principal component (the *X* axis) indicate that states tend to prefer one subset of these weight/ratings or the other, but rarely anything in between.

Cluster 2, in the top center of Figure 1, by contrast, differs from the other two along the second principal component (the *Y* axis) and captures states with weights that emphasize SQ/SS, along with Indicator Dashboard and 1–5 Star ratings. These variables are negatively correlated with Achievement in particular (bottom of *Y* axis), though this negative correlation is weaker than that seen between weights along the first (*X*) principal component. This puts the SQ/SS cluster most clearly in a position distinct from the more common emphasis on student test scores.

What Factors or Variables Help to Explain These Different Responses?

Both the cluster and PCA analyses reveal that states can be organized into three clusters that reflect underlying structures and patterns in their accountability designs. To help explain these structures and patterns, we turn to the independent variables introduced earlier. These include

- *Economic*: Percent of students in poverty and per pupil expenditures. Hypothesis: States with more economic resources—fewer students in poverty and higher per pupil expenditures—are more likely to consider alternative accountability strategies, such as the SQ/SS indicator. In contrast, states with fewer economic resources are more likely to focus on existing and traditional test score indicators.
- *Political*: Policy liberalism index and the political party of the governor in 2017. Hypothesis: More liberal states—high policy liberalism

score and Democratic governors—are more likely to explore alternative indicators, such as SQ/SS. In contrast, conservative states are more likely to focus on existing and traditional test score indicators.

- *Motivational*: Average state NAEP scores in math and reading in fourth and eighth grades in 2017. Hypothesis: Higher achieving states—high average NAEP scores—are more likely to continue a focus on test scores. Lower achieving states are more likely to consider alternative indicator strategies.

In addition to the variables listed above, which are included in a regression analysis, we also consider an *institutional* variable consisting of indicator weights used by states in pre-ESSA accountability systems. Our hypothesis is that states will continue to emphasize indicators used in their pre-ESSA system. As pre-ESSA data are available for only 35 states and the District of Columbia (see Supplemental Appendix C), we use it in a contextual manner to help explain differences among the clusters.

We apply the economic, political, and motivational variables to the cluster and PCA data. In Table 3, we present the average values of the independent variables. This provides a general perspective on the variation across clusters of these key variables. Table 3 also shows which clusters are significantly different across each variable (via *t*-tests); stars between cells indicate that those two clusters differ significantly on that variable, and it is also notable that for every variable shown, Clusters 1 and 3 are significantly different. As with our PCA analysis, the most significant dimension of variation in our data is between Clusters 1 and 3, or equivalently, along the first principal component (*X* dimension in Figure 1), as will be discussed in more detail shortly.

To provide a more nuanced picture of how the independent variables affect indicator designs, we return to the principal component results and regress the two PCA scores on our independent variables to illuminate how those independent variables affect the underlying factors that shape indicator designs. The results of this analysis are in Table 4, which shows both bivariate regressions between individual pairs of variables, as well as ordinary least squares (OLS) multivariate results.

The bivariate analyses in Table 4 (Columns 1 and 2) show significant relations between our independent variables and the principal components that underlie the variation in state accountability choices. Columns 3 and 4 in Table 4 show multiple regression results on the first and second principal components (PC), where each variable's contribution is measured holding all others constant. Notably, unlike the bivariate comparisons, here the partial correlations of the independent variables are distinctly segregated across the two principal components: The first component, PC1, is only associated with

Table 3. Average Values for Independent Variables by Cluster.

Cluster	States	Per pupil expenditures	Students in poverty %	NAEP average	Policy liberalism index	Governor's party
Cluster 1	AL, AR, AZ, FL, IN, MI, MS, NC, NM, OH, OK, PA, TN, TX, UT, VT	US$10,162	21.1	251	−2.96	Republican—14 Democrat—2
		**			**	*
Cluster 2	CA, DC, IL, KY, LA, MD, ND, NV, OR, RI, WV	US$12,722	19.8	248	3.03	Republican—5 Democrat—6
			**	**		
Cluster 3	AK, CO, CT, DE, GA, HI, ID, IA, KS, MA, ME, MN, MO, MT, NE, NH, NJ, NY, SC, *SD*, VA, WA, WI, WY	US$13,090	14.9	254	0.71	Republican—14 Democrat—9 Independent—1
Average		*US$12,092*	*17.9*	*252*	*0*	

Note. All Cluster 1 versus Cluster 3 values for all variables are also significantly different ($p < .05$). NAEP = National Assessment of Educational Progress.
Values above and below stars are significantly different (**$p < .05$. *$p < .10$).

Table 4. Regression Results of Independent Variables with Principal Components.

	Bivariate coefficients		OLS coefficients	
	PC1	PC2	PC1	PC2
Independent Variables	(1)	(2)	(3)	(4)
Per pupil spending	−.237*** (.061)	.018 (.057)	−.181** (.080)	−.091 (.072)
Pct. students in poverty	.144*** (.040)	.011 (.037)	.123** (.059)	−.039 (.053)
Policy liberalism	−.098** (.040)	.072** (.033)	.005 (.048)	.108** (.043)
Mean NAEP	−.079* (.042)	−.053 (.035)	.028 (.054)	−.085* (.048)
Governor's party	.816* (.459)	−.312 (.378)		
Constant			−6.978 (14.446)	23.204* (12.971)
Observations	51	51	51	51
R^2			.326	.186
Adjusted R^2			.267	.115

Note. NAEP = National Assessment of Educational Progress; OLS = ordinary least squares; PC = principal component.
*$p < .1$. **$p < .05$. ***$p < .01$.

economic variables: per pupil spending and percentage of students in poverty.[3] This is the dominant relationship that relates Clusters 1 and 3. By contrast, the second principal component, PC2, aligns with one political variable, the policy liberalism index, as well as the motivational variable, NAEP scores (although the latter association, at $p = .08$, falls below traditional thresholds of statistical significance). Thus, PC1 (and the Cluster 1 versus 3 difference) seems to reflect mainly the independent economic effect, while PC2 (Cluster 2 versus 1 and 3) seems to mainly reflect the independent political effect.

Building from this analysis, Cluster 1 states, on the right side of Figure 1, face significant economic challenges. These states with high PC1 scores have lower per pupil spending and higher student poverty. This is evident in Table 3, with the average values for per pupil spending the lowest and the student poverty rate the highest among the three clusters. Table 4, Column 1, shows bivariate regressions of PC1 scores on state variables and reveals the same pattern: states with a high PC1 score (which tend to be in Cluster 1) have lower spending and higher poverty. These states are most likely to have diminished resources to support public education while also facing a significant challenge with a higher level of student poverty. Table 3 also shows that Cluster 1/high PC1 states are generally conservative-leaning, with a policy liberalism score of −2.96, and 14 of 16 states (88%) having Republican governors during the period when ESSA plans were drafted. But in a multiple regression (Table 4, Column 3), we see that controlling for economic conditions, policy liberalism is not correlated with PC1 scores: the economic variables are the main explanatory factors behind higher PC1 scores and membership in Cluster 1.

The Cluster 1 profile that emerges is a group of states facing economic challenges and opting for a "tough" approach to accountability, primarily in the A–F overall rating system. These are states that stay close to a traditional focus on student test scores, although their emphasis on growth scores indicates a recognition that growth and improvement are important stepping stones for academic achievement. Cluster 1 NAEP scores are only one point below the national average, so low test scores do not appear to be a driving motivation, but there remains a concern for monitoring growth amid the overall preference for student test scores as the major indicators. This weak economic position and focus on test scores are consistent with our earlier hypothesis for the economic variables.

Cluster 3 states are on the other end of this economic dimension. These states on the lower end of PC1 have significantly higher per pupil spending and lower student poverty. This is evident also in Table 3, with the average values for per pupil spending the highest and the student poverty rate the lowest among the three clusters. In addition, this cluster has the highest NAEP

scores, reducing comparative pressure for changes in school practices. As with Cluster 1, the primary effect driving low PC1 scores is economic, with political characteristics showing no conditional relation to PC1 scores. But quite distinct from Cluster 1, states in Cluster 3 are well-resourced and are under the least pressure, relative to states in the other clusters, to make changes in policies and programs.

Politically, in descriptive terms, Cluster 3 is in the moderate range of the spectrum. Along the measures we have used, state governorships in this cluster in 2017 were filled by Republicans in 14 of the 24 states (58%), and the policy liberalism index is a middle-of-the-road 0.71. In combination with the strong economic variables, it is not surprising that states in this cluster have followed a middle path with respect to giving weight to accountability indicators, with a slight preference for High School Achievement and English Learners, along with a strong preference for Graduation. This pattern deviates, in part, from our economic hypothesis. Specifically, the stronger economic position of states in this cluster points less to innovation and more to a continued emphasis on test scores, like in Cluster 1. However, there is a shift away from Growth test scores at the high school level in favor of the Graduation indicator, and the relative emphasis on English Learners also sets this cluster apart. The movement away from Growth test scores is shown as well in the institutional variable (see Supplemental Appendix C) in which Cluster 3 states lower their weight allocation to Growth from the pre-ESSA to their ESSA plans.

Cluster 2 has a very different profile with respect to the independent variables. States in this cluster, located in the top half of Figure 1, are more clearly distinguished along a political dimension. These states with high PC2 scores are positively correlated with the policy liberalism index (Table 4, Column 2). This orientation is evident in Table 3 as well, with the highest policy liberalism rating (3.03) among the three clusters and 6 of 11 governorships (55%) held by Democrats, the highest among the clusters. NAEP also appears to play a role. While not significantly associated with PC2 scores in the bivariate correlations, when all state-level measures are included in the multiple regression (Table 4, Column 4), we see that NAEP scores are negatively associated with high PC2 scores and Cluster 2 membership. Thus, NAEP scores for this cluster are lower than the national average and lower than the other two clusters, adding a possible motivating factor to explore alternative policy approaches.

In this context, Cluster 2's emphasis on SQ/SS measures is not surprising. In general, consistent with our hypothesis on political variables, these are states open to alternative strategies and approaches to accountability design and willing to give these approaches greater weight in their accountability systems. The new SQ/SS indicator provides that opportunity. These states

often include the most common SQ/SS measures of chronic absenteeism and college and career readiness, but may go beyond that to include climate surveys, student on-track measures, and other monitoring strategies.

Furthermore, Cluster 2 states point to the potential power of political orientations—in this case, policy liberalism—to alter pre-existing practices. With the institutional variable (see Supplemental Appendix C), we anticipated that states under ESSA would most likely continue the pattern of indicator weights from their pre-existing accountability system. Cluster 2 states, however, did the contrary and actually increased the weight given to non-test measures. In comparison to the other clusters, Cluster 2 shows the most significant shift in the use of indicators from the pre-ESSA period to ESSA. Specifically, 9 of the 11 states in Cluster 2 had a NCLB waiver, which facilitates the calculation of indicator weights. These 9 states increased their weight allocation for elementary SQ/SS-type indicators from 4% prior to ESSA to 20% under ESSA. At the high school level, the shift was from 12% to 31%. States in this cluster did this by significantly lowering their weight allocation to Achievement (see Supplemental Appendix C).

In several states, this pattern of policy liberalism is re-enforced by legislative changes. In California, for example, state legislation passed in 2013 requires all school districts to develop a Local Control Accountability Plan that meets a variety of state priorities, several of which are best monitored through SQ/SS measures, such as college and career readiness. In Maryland, the Protect Our Schools Act of 2017 restricts test scores and other academic indicators to no more than 65% of a school's composite accountability score. By state statute, there must be at least three SQ/SS measures, including a school climate survey.

An interesting overall picture emerges. A political dimension appears to be a key driver for innovative measures such as SQ/SS, while an economic dimension aligns with a more traditional focus on student testing. Specifically, the states in Cluster 2 are leading the experiment with SQ/SS measures, and they are doing so with the sharpest break from their pre-ESSA configuration of accountability indicators. A key driver for this is a liberal pattern of policymaking. States in this cluster are taking a leadership role in exploring alternative accountability measures and metrics *outside* student test scores. In contrast, Clusters 1 and 3 have a stronger association with economic factors and show less inclination to deviate from past practice. Cluster 1 shifts to a greater emphasis on growth measures and the stricter A–F rating system, and Cluster 3 demonstrates some shifts in indicator preference and a more flexible rating system, but both remain more closely aligned to existing accountability practices.

What Are the Consequences for Education Politics?

Accountability systems are fundamentally about the creation and use of information. That is not a trivial matter. Information can privilege certain perspectives and prompt the shift of resources to particular groups and interests. As Dorothy Anagnostopoulos and her colleagues (2013) note in their study of test-based accountability, this "informatic power" can "reshape the ways in which Americans practice, organize, participate in, and even think about the nation's public schools " (p. 2). To borrow from the policy design and feedback literature, accountability information—indicators and overall ratings—has important interpretive effects on how we understand and conceptualize policy issues as well as resource effects on how financial, personnel, and other resources are allocated (Campbell, 2012).

The exercise of "informatic power" is an important vantage point on politics. It involves the contest among different interests to give preference to particular perspectives on how accountability systems should be designed and implemented. Giving greater relative weight to growth scores over achievement, for example, shifts not only the attention of policymakers, educators, parents, and the general public but can also shift the allocation of resources within a school system. Similarly, labeling a school with the grade of "C" as opposed to identifying weaknesses across attendance and graduation metrics in a dashboard can shape the perception of a school and how resources are distributed.

In this context, our cluster and PCA analyses point to several variations in education politics across the American states. Cluster 1 states, or more generally, those with higher PC1 scores, align with a more conservative politics and economic challenges, a "tough" approach to accountability, and an emphasis on test scores. With an A–F Grading system, used by most states in this cluster, policymakers, who are typically non-educators, opt for a direct and unequivocal message to educators and community members on a school's performance. Information in this context serves primarily as a tool for non-educators. A traditional emphasis on standardized tests remains prominent, with a nod to growth scores as a way to balance the long-standing use of achievement scores. The economic challenges faced by states in this cluster undergird this more conservative political orientation.

Cluster 2, or higher PC2 scores, points to a different kind of politics. With a more liberal orientation, these states are most likely to explore alternative accountability designs, as evident in their emphasis on SQ/SS measures. These states take most seriously the push to incorporate measures and metrics beyond test scores. Climate surveys, school quality reviews, and similar types of information can serve a formative role in helping educators improve

the environment for teaching and learning. Information is an instrument to support change and continuous improvement. In contrast to the summative, external emphasis of information in Cluster 1 accountability, Cluster 2 points to a politics more attuned with a school-based perspective on school improvement. In comparison to states in other clusters, these states are most willing to move beyond their pre-ESSA accountability pattern. Interestingly, this cluster includes states with overall rating systems—1–5 Stars and Indicator Dashboard—that package information in quite different ways. That is, a 1–5 Stars approach uses a single label for a school, but the Indicator Dashboard specifically avoids that approach, opting instead for a multiple metrics design. A common thread for both, however, is a liberal policy orientation.

A recent exchange between former Florida Governor Jeb Bush, from a Cluster 1 state, and Michael Kirst, past president of the California Board of Education, a Cluster 2 state, provides an example of these different political orientations and the power of information. Bush, an early advocate of the A–F approach, argues that letter grades provide parents with a "clear and concise measure of school performance" while "focusing educators on the goal of maximizing academic achievement." From Bush's perspective, a dashboard, as used by California, is a "fog machine" with no transparency. Kirst fires back, critical of the negative message that comes from letter grades, arguing that they are "not only uninformative, they are inaccurate when it comes to identifying low-performing schools." A dashboard is a more "holistic" and "comprehensive picture of a school's successes and challenges" (Education Next, 2017, pp. 57–62). This exchange portends a political clash that is likely to grow in the future.

Cluster 3—states with lower PC1 and lower PC2 scores—offers a more variable environment in which politics may differ among states within the cluster. This is most evident in the Performance Level rating approach used by many states in Cluster 3. In this rating system, accountability information in assessing school performance is tailored to the different performance standards and expectations established in each state. The political dialogue that follows is shaped by this state-specific pattern of interests and perspectives. Also, the relative emphasis in these states on indicators for English Learners and Graduation highlight these types of information as important parts of the political dialogue. Cluster 3 states, then, provide potential for a broader array of state and school-specific political debate and exchange.

Conclusion

As the pendulum swings back to the states, the design of educational accountability systems, as presented in state ESSA plans, is taking states in

several different directions. Our analysis of indicator weights and overall rating systems provides an important window on this process. Some states are making minor adaptations and adjustments to their indicator weights, while others are engaging in more significant and robust redesign efforts. From our analysis, three clusters of states emerge in which policy stasis or change are driven by a variety of underlying political, economic, motivational, and institutional factors. Of particular note is an economic dimension that underlies two clusters and is associated with a state focus on test-based indicators, distinct from a separate political dimension that highlights a third cluster with a liberal orientation and an emphasis on innovation in the form of the SQ/SS indicator.

Accountability systems under ESSA will continue to evolve as states adjust and adapt to a changing environment. To be certain, some states will be more engaged and proactive in this process than others. Yet, ESSA does mark an important shift. The federal-to-state "collision" that Paul Manna (2011) described under NCLB has transitioned under ESSA to a reformulated partnership, albeit within various federal "guard rails," in which states are playing a major role in shaping education policy. Most importantly, in assuming this role, they are following *different* policy paths that will increasingly reflect their own interests and concerns, including their own political perspectives and experiences. The often-noted political polarization of American politics and the states themselves is likely to be reflected in evolving accountability systems. It is in this environment that policy design becomes so important as new state accountability systems help shape the future path for an increasingly diverse world of education politics and policy.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

John Portz https://orcid.org/0000-0003-4365-7697

Supplemental Material

Supplemental material for this article is available online.

Notes

1. A note on terminology: *Indicator* is the term used broadly in reference to the general construct of information to assess whether a goal has been met. *Measure* is the way in which an indicator is assessed. *Metric* refers to how a measure is scored in an accountability system. As an example, academic proficiency is an indicator; a state achievement test is the measure; and the student's score on the test is the metric. At times, there is overlap between these categories, but the distinctions are helpful in a discussion of accountability design (Buckley, 2017).
2. The A–F and 1–5 Stars rating systems are the terms used by the states. This category of Performance Level, as well as the next two categories—Support Levels and Indicator Dashboards—are titles developed largely by the authors, based on an analysis of each state's ESSA plan.
3. We also examined state median income, but median income is highly correlated with poverty rate and is uncorrelated with either principal component when controlling for poverty and thus was omitted from the main model. Similarly, the governor's party is not included in the multiple regression because it is strongly correlated with policy liberalism but is not significant when controlling for policy liberalism. Results are weaker but not substantively changed when using median income or governor's party.

References

Achieve. (2018). *College and career readiness*. https://www.achieve.org/college-and-career-readiness

Anagnostopoulos, D., Rutledge, S. A., & Jacobsen, R. (2013). Mapping the information infrastructure of accountability. In D. Anagnostopoulos, S. A. Rutledge, & R. Jacobsen (Eds.), *The infrastructure of accountability: Data use and the transformation of American education* (pp. 1–20). Harvard Education Press.

Baumgartner, F. R., & Jones, B. D. (2015). *The politics of information: Problem definition and the course of public policy in America*. University of Chicago Press.

Berry, F. S., & Berry, W. D. (2018). Innovation and diffusion models in policy research. In C. M. Weible & P. A. Sabatier (Eds.), *Theories of the policy process* (4th ed., pp. 253–300). Westview Press.

Boehmke, F. J., & Skinner, P. (2012). State policy innovativeness revisited. *State Politics and Policy Quarterly*, *12*(3), 303–329.

Boushey, G. (2010). *Policy diffusion dynamics in America*. Cambridge University Press.

Brown, C. G. (2017). From ESEA to ESSA: Progress or regress? In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act (ESSA): What it means for schools, systems, and states* (pp. 153–169). Harvard Education Press.

Buckley, K. (2017). *Innovative approaches and measurement considerations for the selection of the school quality and student success indicator under ESSA*. Council of Chief State School Officers.

Campbell, A. L. (2012). Policy makes mass politics. *Annual Review of Political Science*, *15*, 333–351.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Center on Education Policy. (2012a). *Accountability issues to watch under NCLB waivers*.

Center on Education Policy. (2012b). *What impact will NCLB waivers have on the consistency, complexity and transparency of state accountability systems?*

Center on Education Policy. (2017). *Planning for progress: States reflect on year one implementation of ESSA*.

Council of Chief State School Officers. (2011). *Roadmap for next-generation state accountability systems*.

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Poldolsky, A., & Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. Learning Policy Institute.

Darling-Hammond, L., & Hill, P. T. (2015). *Accountability and the federal role: A third way on ESEA*. Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, *22*(86), 1–38. http://doi.org/10.14507/epaa.v22n86.2014

Data Quality Campaign. (2019). *Growth data: It matters, and it's complicated*. https://dataqualitycampaign.org/resource/growth-data-it-matters-and-its-complicated/

D'Brot, J. (2017). *Considerations for including growth in ESSA state accountability systems*. Council of Chief State School Officers, National Center for the Improvement of Educational Assessment.

Education Commission of the States. (2018). *50-State comparison: States' school accountability systems*. https://www.ecs.org/50-state-comparison-states-school-accountability-systems/

Education Next. (2017). How should states design their accountability systems? *Winter*, *17*(1), 55–62.

Hamilton, L. S., Schwartz, H. L., Strecher, B. M., & Steele, J. L. (2013). Improving accountability through expanded measures of performance. *Journal of Educational Administration*, *51*, 453–475.

Hess, F. M., & Eden, M. (2017). *The Every Student Succeeds Act (ESSA): What it means for schools, systems, and states*. Harvard Education Press.

Idaho ESSA Plan. (2018). *Idaho's consolidated state plan*. https://www.sde.idaho.gov/topics/consolidated-plan/files/Idaho-Consolidated-State-Plan-2019-Amendment.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). Springer.

Jochim, A. (2017). ESSA and state policy: What's next for education policy? In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act (ESSA): What it means for schools, systems, and states* (pp. 121–136). Harvard Education Press.

Kaput, K. (2018). *Briefing memo: Summaries of states' ESSA school quality/student success measures*. Education Evolving.

Klein, A. (2017). How ESSA passed: The inside scoop. In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act (ESSA): What it means for schools, systems, and states* (pp. 43–58). Harvard Education Press.

Klein, A. (2019, January 22). How are states measuring college-and-career readiness? It's a hodge-podge. *Education Week*. https://www.edweek.org/ew/articles/2019/01/23/states-using-post-high-school-readiness-as-essa.html

Kostyo, S., Cardichon, J., & Darling-Hammond, L. (2018). *Making ESSA's equity promise real: State strategies to close the opportunity gap*. Learning Policy Institute.

Lyons, S., D'Brot, J., & Landl, E. (2017). *State systems of identification and support under ESSA: A focus on designing and revising systems of school identification*. Council of Chief State School Officers.

Manna, P. (2011). *Collision course: Federal education policy meets state and local realities*. CQ Press.

Martin, C., Sargrad, S., & Batel, S. (2016). *Making the grade: A 50-state analysis of school accountability systems*. Center for American Progress.

McGuinn, P. J. (2006). *No child left behind and the transformation of federal education policy, 1965-2005*. University Press of Kansas.

McNeil, M. (2014, April 10). Many states left key NCLB flexibility on the table. *Education Week*. https://www.edweek.org/ew/articles/2014/04/10/28multiple.h33.html

Nice, D. (1994). *Policy innovation in state government*. Iowa State University Press.

Pierson, P. (2004). *Politics in time: History, institutions, and social analysis*. Princeton University Press.

Rafa, A. (2017, June). *Chronic absenteeism: A key indicator of student success*. Education Commission of the States.

Saultz, A., Fusarelli, L. D., & McEachin, A. (2017). The Every Student Succeeds Act, the decline and of the federal role in education policy, and the curbing of executive authority. *Publius: The Journal of Federalism*, *47*(3), 426–444.

Schneider, J. (2017). *Beyond test scores: A better way to measure school quality*. Harvard University Press.

Shober, A. F. (2017). ESSA and state capacity: Can the states take accountability seriously? In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act (ESSA): What it means for schools, systems, and states* (pp. 107–120). Harvard Education Press.

U.S. DOE. (2013). *ESEA flexibility*. http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Utah ESSA Plan. (2017). *Revised state template for the consolidated state plan*. https://www.schools.utah.gov/file/5de70d64-5f04-4d46-bf65-49e9ed4dce2e

West, M. R. (2017). The case for ESSA: A proper balance. In F. M. Hess & M. Eden (Eds.), *The Every Student Succeeds Act (ESSA): What it means for schools, systems, and states* (pp. 75–90). Harvard Education Press.

Author Biography

John Portz is Professor of Political Science at Northeastern University. His areas of research interest include public policy and federalism, with a current focus on educational accountability and governance. He also serves on the school board in his home community of Watertown, MA.

Nicholas Beauchamp is Assistant Professor of Political Science and a core faculty member at the Network Science Institute at Northeastern University. His research applies techniques from natural language processing, machine learning, and network science to better understand political communication and policies.