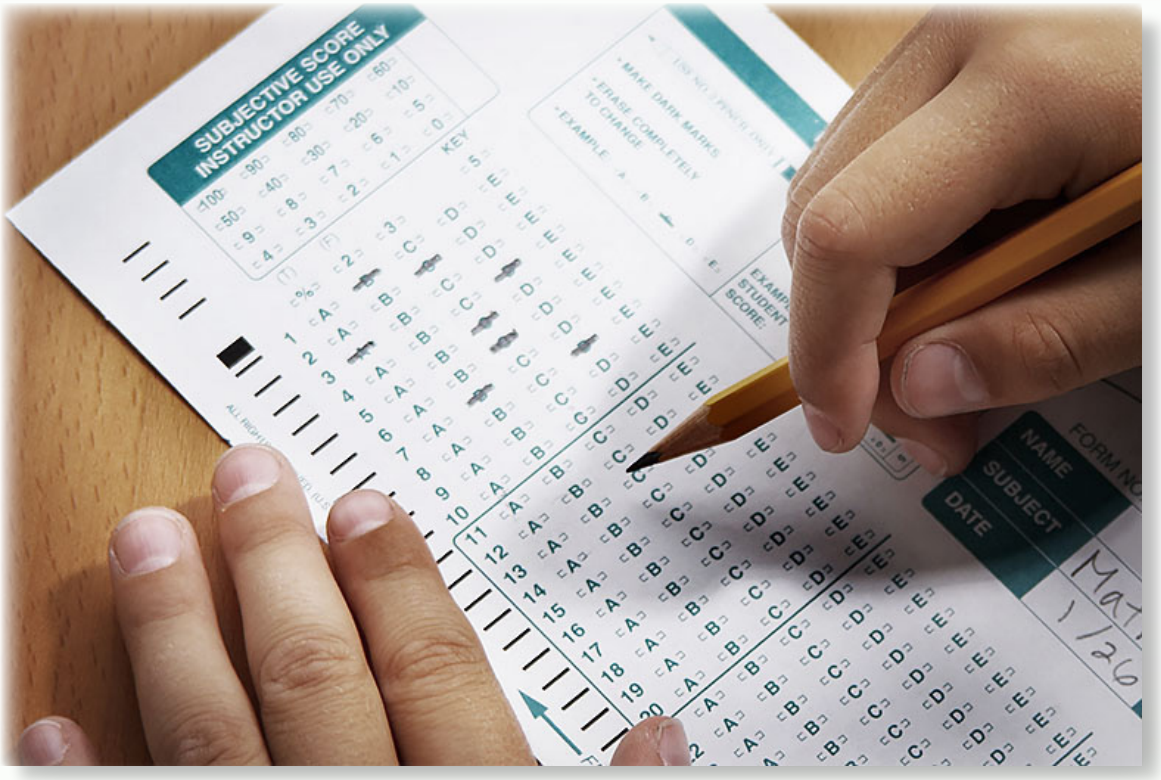
State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction



Kevin Close, Audrey Amrein-Beardsley, and Clarin Collins
Arizona State University

June 2018

National Education Policy Center

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
(802) 383-0058
nepc.colorado.edu

Acknowledgements

NEPC Staff

Kevin Welner
Project Director

William Mathis
Managing Director

Patricia Hinchey
Academic Editor

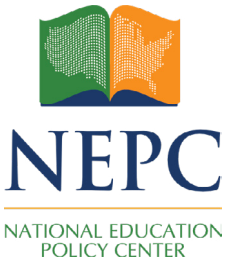
Alex Molnar
Publications Director

Suggested Citation: Close, K., Amrein-Beardsley, A., & Collins, C. (2018). *State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/state-assessment.

Funding: This policy brief was made possible in part by funding from the Great Lakes Center for Educational Research and Practice.

Peer Review: *State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction* was blind peer-reviewed.

This publication is provided free of cost to NEPC's readers, who may make non-commercial use of it as long as NEPC and its author(s) are credited as the source. For inquiries about commercial use, please contact NEPC at nepc@colorado.edu.



State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction

Kevin Close, Audrey Amrein-Beardsley, and Clarin Collins
Arizona State University

June 2018

I. Executive Summary

Federally mandated standardized testing (i.e., in core subject areas and certain grade levels), as an element of educational accountability, began in 2002 with the No Child Left Behind Act (NCLB). The Act mandated that states, among other requirements, develop or adopt tests in reading and mathematics and administer them annually to students in grades 3-8 and once in high school. These large-scale tests were to serve as a basis for measuring student achievement and to hold schools and districts accountable. Eventually, they also served as a basis for measuring teacher effectiveness, a strategy promoted through the subsequent Race to the Top (RttT) initiative and through requirements for schools seeking waivers from NCLB requirements. In short, large-scale assessments have come to serve as one of the foundations of accountability-based systems and policies not only for districts, schools and students, but for teachers as well.

However, both before and after NCLB and RttT the academic community, policymakers, administrators and teachers noted weaknesses in these policies and systems. Researchers in particular investigated and subsequently questioned the practice of tying high-stakes consequences to scores on such large-scale assessments, especially at the student and teacher levels. As a result of these critiques and other concerns, Congress passed the 2015 Every Student Succeeds Act (ESSA), which reduced federal oversight and gave states more control over their state assessment and accountability systems, weakening the federal influence on teacher evaluation systems promoted in RttT and in waiver requirements.

This brief offers a thematic analysis of state-level assessments in ESSA plans from each state and the District of Columbia. It also includes results of the authors' detailed survey about state teacher evaluation systems. The survey was completed by department of education

personnel from 34 states and the District of Columbia. Analyses of ESSA plans and survey information indicate that, in general, states continue to use the same large-scale student tests in place before ESSA, and they continue to give those test results a role in evaluations of teacher effectiveness. However, greater local control has led to some encouraging signs of change. These include: efforts to redefine student growth as something other than growth in test scores; movement toward multiple assessment tools, including student learning objectives (SLOs); fewer states emphasizing value-added assessments in teacher evaluations; and a move away from high-stakes consequences and toward formative rather than summative assessments.

As a whole, this brief should be read as a "state-of-the-states" report on what has happened to student and teacher evaluation systems since the passage of ESSA.

Recommendations

Based on findings from the analyses and survey information, it is recommended that state policymakers take the following five steps:

1. Take advantage of decreased federal control by formulating revised assessment policies informed by the viewpoints of as many stakeholders as feasible. Such informed revision can help remedy earlier weaknesses, promote effective implementation, stress correct interpretation, and yield formative information.

2. Ensure that teacher evaluation systems rely on a balanced system of multiple measures, without disproportionate weight assigned to any one measure.

3. Emphasize data useful as formative feedback in state systems, so that specific weaknesses in student learning can be identified, targeted and used to inform teachers' professional development.

4. Mandate ongoing research and evaluation of state assessment systems and ensure that adequate resources are provided to support evaluation.

5. Set goals for reducing proficiency gaps and outline procedures for developing strategies to effectively reduce gaps once they have been identified. Finding which interventions work best to reduce proficiency gaps will take some experimentation. Have a procedure in place to develop these interventions.

NEPC
NATIONAL EDUCATION POLICY CENTER

State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction

Kevin Close, Audrey Amrein-Beardsley, and Clarin Collins
Arizona State University

June 2018

II. Introduction

The No Child Left Behind (NCLB) Act of 2002 mandated that states develop or adopt tests to measure student achievement in reading and mathematics, administering them annually in grades 3-8 and once in high school (with science testing to be phased in later). This marked a new era in state-level testing and assessment-based accountability policies and systems. NCLB further required states to develop accountability plans that tracked statewide progress toward the goal of 100% student proficiency for students overall and for designated subgroups. And, schools had to meet goals for adequate yearly progress (AYP) as they worked toward the 100% proficiency goal for students and subgroups. Schools not meeting AYP became subject to state sanctions ranging from having to provide supplemental education services to being restructured.

After student-testing and proficiency goals were established, federal focus shifted from accountability for student performance to accountability for teacher performance. A federal grant competition announced in 2009, Race to the Top (RttT), provided incentives for states to incorporate measures of student achievement and growth into systemic assessments of teacher effectiveness. Additionally, when NCLB was not promptly reauthorized after it expired, in 2012 the Obama administration began offering waivers to states not achieving proficiency goals. To obtain a waiver, states had to adopt certain reforms—including incorporating student achievement into teacher evaluation systems.

Some states measured teacher impact on students' achievement by measuring student growth; that is, they focused on changes in test scores over time. This metric has been called teacher-level growth, or "value-added" models (VAMs). However, using test scores over time and attributing all growth or lack of growth to teachers caused much debate and conflict. Some districts reported difficulty retaining teachers after the implementation of such

strong, teacher- and test score- focused accountability systems. Some teachers, and teacher unions, filed lawsuits challenging such systems, and especially the consequential decisions being made in some states based on such measures. Many in the academic community wrote articles primarily about empirical issues, but also about the same pragmatic issues being noted in the field. In addition, leading educational/professional organizations such as the American Educational Research Association (AERA) and American Statistical Association (ASA) issued official statements about problems with the validity of such measures as well as their subsequent uses. And, some states and districts outright opposed federal policies by rejecting the waivers, refusing to accept dictates about how to evaluate local teachers given their unique demographic, political, or geographical considerations.

On January 1, 2016, then-President Obama signed the Every Student Succeeds Act (ESSA), which reined in some of the strict federal mandates and oversight regarding states' assessment and teacher evaluation systems. ESSA gave states and districts more control over the design and implementation of assessment-based policies and systems, and it also allowed them more leniency in interpreting concepts like "including as a significant factor, data on student growth for all students." In short, ESSA has presented an opportunity for states to change their educator evaluation systems.

Given the number of years and amount of significant financial and human resources that states invested in developing and implementing federally required accountability and teacher evaluation plans, and given known problems stemming from federal mandates, this brief asks to what extent states have taken advantage of new flexibility and revised their assessment and teacher evaluation plans. The answer to that question is based on analyses of current state plans and survey responses from personnel in departments of education. More specifically, sources of data included: a) ESSA plans from every state and the District of Columbia, henceforth called the 51 plans, and b) a 30-minute computer-based survey the research team designed, developed, validated, and then administered to state department of education personnel. The following section addresses the details of the research design.

Methodology

Over one year, the researchers collected the 51 plans from education department personnel via a survey (n = 20) and from the US Department of Education website (n = 31). The survey also asked education department personnel to provide detailed information about current teacher evaluation systems, including questions about perceived strength and weaknesses, value-added models, and consequences tied to the evaluation systems. Of the 51 units contacted, department personnel in 34 (≈67) responded to the online survey; another three (≈6%) answered survey questions via a phone interview. In two cases (≈4%), personnel did not answer the survey questions but did refer the researchers to online resources. For the other 12 units (≈24%), the researchers filled in missing information using publicly-available ESSA plans or state websites. Department personnel who responded to the survey typically worked as teacher effectiveness directors or coordinators. Occasionally, the directors or coordinators recommended other respondents (leadership consultants or educator effective-

ness specialists, for example) they felt could better answer the questions.

Initially, the researchers read each plan focusing only on sections that related to state-level assessments and teacher evaluation systems. Within these sections, researchers first read to become familiar with the data, then generated initial ideas, and subsequently formulated overall themes based on the earlier readings and analyses. That is, the research team systematically looked for trends in state-level assessments within the ESSA plans. For the sake of interpretability and presentation, the researchers also noted exemplary cases for each trend.

Most survey questions yielded straightforward frequency counts (for example, 15 of 51 participants reported that their state encouraged statewide use of growth models or VAMs). For open response questions, the researchers collapsed and quantified responses to note common themes.

As a whole, this brief should be read as a "state-of-the-states" report on what is happening post-ESSA in state-level assessments, with an emphasis on teacher evaluation systems. Though other publications examine ESSA state plans as well, this brief targets state-level assessments and teacher evaluation systems to provide a simple overview. Additionally, this brief provides hypotheses about why some types of teacher evaluation systems may persist, in some cases despite the research evidence against them.

III. Review of the Relevant Literature

As noted, the early 2000s marked a new era in educational accountability policies, with federal policies increasingly promoting state-, district-, school-, teacher-, and student-level test or assessment systems that held educators accountable for student achievement results. This shift constituted a change in the testing landscape, originally characterized by NCLB's mandate for states to devise standards and to use large scale assessments to track student mastery of those standards. The theory of change was that by holding districts, schools, teachers, and students accountable for meeting higher standards, as measured by student performance on such assessments, administrators would supervise America's public schools better, teachers would teach better, and as a result, students would learn and achieve more, particularly in America's lowest performing schools.

However, most researchers now agree that NCLB was, overall, a failed federal policy. Current consensus in this area is that NCLB did not meet its intended effects (100% student mastery of higher standards by 2014). Rather, NCLB caused unintended effects (including artificial inflation of test results, teaching to the test, cheating, and other system gaming techniques), and its unintended effects outweighed the other positive effects noted (an increased focus on measuring and monitoring the gap between marginalized and non-marginalized student populations). It should be noted, however, that not all researchers agree.

Perhaps acknowledging the issues with NCLB, the federal government soon used federal funds again to entice states and districts to move in new directions. Specifically, they were

encouraged to adopt new and improved tests (those developed by the Partnership for Assessment of Readiness for College and Careers [PARCC] or Smarter Balanced Assessment Consortium [SBAC]) and by using new and improved systems and policies—new practices but based on the same change theory. Test-based calculations of student growth over time using the NCLB-mandated tests (value-added measures, or VAMs) ultimately replaced the adequate yearly progress (AYP) measurements. That is, the federal government began advocating the use of test results not only to measure students' growth in learning over time, but also to measure teachers' causal impacts on that growth. For example, Race to the Top (RttT) was a grant incentive used to move states in those directions.

Soon after RttT was underway, in 2014, 40 states and the District of Columbia (80% of the 51 units reviewed in this brief) were using, piloting, or developing some type of growth model or VAM, as incentivized by the federal government. The tests required under NCLB were being used across the VAMs in place at the time as a base for measuring teacher-level value-added. The most common open-source VAM model was student growth percentiles (SGP), with multiple states adopting or endorsing it (Arizona, Colorado, Georgia, Massachusetts, Washington); the most common proprietary model was the Education Value-Added Assessment System (EVAAS), with five states adopting it (North Carolina, Ohio, Pennsylvania, South Carolina, and Tennessee). The most common high-stakes consequences being attached to VAM results involved teacher tenure, termination, and compensation (merit pay, for example).

While the teacher evaluation systems being adopted and implemented at this time included multiple other indicators or measures of effectiveness (for example, classroom observations of teachers), the primary focus across states was on an objective, assessment-based component to "meaningfully differentiate [teacher] performance…including as a significant factor, data on student growth [in achievement over time] for all students." This strategy was written into federal policy and widely implemented, although some states (Florida, Louisiana, Nevada, New Mexico, New York, Tennessee, Texas) weighted student growth much more heavily in their systems than other states did (California, Connecticut, Vermont, Washington, Wisconsin). Because of NCLB requirements, the achievement tests needed to provide a base for VAM systems were already being paid for and in place or being developed (PARCC and SBAC).

In the simplest of terms, statisticians use growth models or VAMs to measure the predicted and then actual "value" a teacher adds to student achievement on such tests from year to year. Modelers typically do this by predicting and then measuring student growth over time on large-scale assessments, controlling statistically for confounding variables such as students' prior test scores and other student' and school-level variables, and then aggregating growth at the teacher-level. Controls vary by model. What many promoting such assessment-based systems are still quick to forget, or dismiss, is that there is very little empirical evidence to support attaching high-stakes consequences to resulting measures of teacher effectiveness.

In addition, what was becoming increasingly evident by the time that ESSA was passed, yet

also simultaneously increasingly marginalized in educational policy, was that VAM models are notoriously (1) unreliable (for example, a teacher classified as adding value has a 25-50% chance to be classified as subtracting value the following year); (2) invalid (that is, very limited evidence that teachers who post high growth or value-added scores are effective using at least one other correlated criterion); (3) nontransparent (teachers and administrators often do not understand the models being used to evaluate them); (4) unfair (only teachers of mathematics and language arts with pre- and post-test data are being held accountable using these systems); and (5), fraught with measurement errors (for example, inordinate amounts of missing data, variables that cannot be controlled, variance caused by the non-random placement of students into classrooms, issues caused by non-traditional and non-insular classrooms). Moreover, they were being used inappropriately to make consequential decisions (including teacher-level professional development, promotion, probation, tenure, merit pay, termination), especially in some states where high-stakes consequences resulted in court challenges to the state systems (Florida, Louisiana, Nevada, New Mexico, New York, Tennessee, and Texas), and their unintended consequences were often going unrecognized. As best summarized by Moore Johnson (2015), unintended consequences included, but were not limited to: (1) teachers being more likely to "literally or figuratively 'close their classroom door' and revert to working alone…[which]…affect[s] current collaboration and shared responsibility for school improvement" (p. 120); (2) teachers being "[driven]…away from the schools that need them most and, in the extreme, causing them to leave [or to not (re)enter] the profession" (p. 121); and (3) teachers avoiding teaching high-needs students who might be more likely to hinder their value-added results, "seek[ing] safer [grade level, subject area, classroom, or school] assignments, where they can avoid the risk of low VAMS scores" (p. 120), a situation leaving "some of the most challenging teaching assignments… difficult to fill and likely…subject to repeated [teacher] turnover" (p. 120).

What is also becoming more evident in the recent literature is that the observational systems used for summative evaluation, a common part of contemporary teacher evaluation systems, are now confronting their own sets of empirical issues. Such issues include, importantly, how output from observational systems might also be biased by such factors as the type of student a teacher works with, the teacher's gender, and the like. The same sorts of potential bias seem to hold true with student surveys, whether used to evaluate teachers in Pre-K or higher education.

The new freedom that ESSA affords means that states could be moving away from such high-stakes and assessment-based accountability models. This brief aims to uncover whether states are actually taking advantage of the most recent, more flexible policy and moving in new directions, away from such models.

IV. Recent Developments and Analysis

NCLB imposed statewide accountability systems including AYP, and RttT provided incentives for states to track students' achievement longitudinally and to adopt common academic standards allowing for cross-state comparisons. For many states, this meant joining large

state consortiums and sharing standards and tests. However, ESSA allows each state to set its own goals within the federal framework, an opportunity not allowed before. While the federal framework still requires measurements of AYP for subgroups, states select intervention plans for a subgroup failing to meet progress goals.

The distinction may seem small, but personnel in departments of education responding to our survey reported that they allowed greater control for stakeholders (that is, less top-down department of education control) by including more stakeholders in the process of developing state-level assessment plans. Though ESSA requires states to include stakeholders in the process, the survey results provide some insight into how that requirement is translating to practice. Under the earlier and arguably more heavy-handed policies and incentives of NCLB and RttT, such stakeholder meetings would have been more difficult. Department personnel personal reported that ESSA has allowed for the creation of more focused evaluation systems that better reflect the desires of stakeholders, moving away from what were sometimes termed two-headed systems: one that appeased federal requirements and one that reflected the desires of the state (and sometimes its stakeholders). No longer hamstrung by NCLB and RttT mandates and incentives, states are, at minimum, showing some encouraging signs of change.

State-Level Assessment Trends

There are three trends of note in some states' assessment trends.

Trend 1: Little to no change regarding the quantitative assessments used to assess student proficiency—yet.

ESSA still mandates achievement tests for grades 3-8 and an achievement test at least once in high school. Although ESSA also calls for broader measures of student performance, including at least one non-cognitive measure, state-level assessment systems still tend to heavily rely on the earlier tests already in place. These include state-specific tests (for example, California's Construction and Skilled Trades, or CAST, tests aligned with Next Generation Science Standards) as well as well-established nationwide tests (for example, Arkansas uses ACT Aspire for grades 3-10, and New Jersey uses PARCC tests to meet federal testing requirements).

States in the process of including such other types of proficiency as post-secondary readiness in their assessment systems often have ESSA plans that lack good descriptions or that simply are not yet fully developed. For example, Iowa expressly plans to measure post-secondary readiness, but its plans also include convening a workgroup to figure out just how to measure it.

Trend 2: State-level assessments of school success exhibit new indicators that go

beyond subject matter/academic test scores, although some states are still grading schools using A-F letter systems.

Despite changes to lessen federal requirements and allow greater state control, current systems of school-level assessment still reflect ideas from past federal mandates. For example, 44 states and the District of Columbia (≈88%, 45/51) still give schools some sort of overall score as a summative, accountability-based evaluation. Though states have not been required to report such summative scores, they have tended to do so because they were required to report school scores publicly. Fourteen of these states (≈27%, 14/51) still give A-F grades, and another 12 (≈24%, 12/51) give summative scores that include comparable performance categories. For example, Iowa still presents a school-level report card, but instead of receiving A-F, schools receive one of the following ratings: Exceptional, High-Performing, Commendable, Acceptable, Needs Improvement, and Priority. The other 19 states (≈37%, 19/51) give a summative rating based on an index score (often 1-100), a tier system, or a star system. Massachusettes, for example, according to their ESSA state plans, has six tiers of schools; Alaska rates schools on a 100-point index scale; and Nevada gives schools a rating from one to five stars.

Only six states (California, Idaho, Oregon, North Dakota, Pennsylvania, and Virginia) (≈12%, 6/51), on the other hand, opted to give no assigned performance category to schools, providing instead a menu of information meant to reflect multiple measures in addition to student achievement. Pennsylvania, for example, opted to present information on chronic absenteeism, student growth, and access to well-rounded and advanced courses, among other measures. In fact, chronic absenteeism and college readiness dominate as major factors in school-level accountability, with at least 35 states mentioning at least one of the two factors. While measures such as chronic absenteeism and college readiness were collected before the passage of ESSA, and therefore collecting these measures does not appear to be a major change, it is notable that states have been making a more holistic view of school performance more readily available.

Trend 3: Using proficiency scores to advance equity is a key narrative, but targets and timelines vary.

The ESSA-based requirement that states still report subgroup scores guided many states to focus on subgroup achievement gaps. Like NCLB, ESSA requires that states include historically underserved subgroups in accountability measurements and decisions, and that they also ensure intervention if certain groups are performing poorly. Differences appear in how states are setting their proficiency goals, how they are establishing their subgroups, and how they are intervening when a subgroup fails to reach proficiency goals. Whereas NCLB proficiency goals were mandated by the federal government, ESSA proficiency goals can be decided by the states. For example, Kansas set a goal of 75% proficiency for all students by 2029-2030. Georgia set a goal of improving proficiency rates by 3% per year for an extended period of time. As is evident from these examples, timing and percentages differ.

In terms of subgroups, ESSA actually tightens control, requiring historically underserved subgroups including racial minorities, English-language learners, those in poverty, and those in special education. Some states over time began to combine subgroups into super subgroups. ESSA counteracts that movement. Most states have a clear plan for which groups to measure, again using large-scale standardized assessments.

Most states also have a plan for how to compare group-level proficiency on those tests to make gaps in proficiency evident. However, state plans do not explicitly address how states might go about reducing proficiency gaps that may continue to persist. Again, under ESSA, states decide their own intervention plans, provided they are evidence-based, to reduce proficiency gaps. However, state ESSA plans reveal that many states do not have clear interventions in place.

Teacher Evaluation Trends

While policymakers, academics, and others may continue to tout ESSA as an ambitious opportunity to reform states' teacher evaluation systems, it seems that most of states have not made substantive changes to policies of the recent past. Some states (Louisiana, Delaware, Florida) are keeping the same teacher evaluation systems in place—perhaps, as noted above, understandably, given the human and monetary resources and monetary already invested in formerly mandated designs. Other states (Texas, Maine), however, are now allowing local education authorities to select from menus or curated lists of recommended teacher evaluation models or measurement components (for example, VAM-based systems, observational rubrics, survey instruments)—all of which also include the teacher evaluation models already in place. The result seems to be that across the board, states are requiring, endorsing, or recommending teacher evaluation systems that are the same or slightly different versions of the previously required systems.

Despite no large shift in direction, there are three trends of note evident in some states' most recent teacher evaluation plans.

Trend 1: The role of growth models or VAMs for teacher evaluation purposes is slowly changing.

The number of states using statewide VAMs has decreased from 42% to 30% since 2014.

As shown in Figure 1, currently 15 states explicitly impose or encourage statewide use of growth models or VAMs (≈30%, 15/51), 21 states explicitly do not use or encourage statewide use of growth models or VAMs (≈42%, 21/51), and 15 states and the District of Columbia (≈30%, 15/51) report the use of "other" approaches when asked if they encourage statewide use of growth models or VAMs. "Other" in this case often means that the states truly offer local control by allowing alternative student growth measurements; alternatives might include VAM but without state encouraging or prescribing a specific VAM model. For example,

Texas falls in the "other" category in that it now emphasizes local control and allows student growth to be measured in one of four ways: 1) student learning objectives (SLOs), 2) portfolios, 3) district-level pre- and post-tests, and 4) VAMs for teachers in state-tested subjects.

ESSA also allows more leniency in how states define student growth. Whereas student growth formerly meant the growth in a student test score over time, aggregated at the teacher level to indicate teachers' impacts over time, the definition of growth under ESSA is much more open to interpretation. Some states simply redefined it. For example, while Connecticut still requires 45% of the teacher evaluation system to be based on student growth indicators, it recently prohibited student growth as measure by the state test from being used in the final summative teacher evaluation. Instead, student growth is measured through a combination of standardized measures (for example, test scores from AP-exams, the SAT-9, the DRA, or others) and non-standardized measures (such as portfolios of student work, performance rated against a rubric). Though not an extreme departure from the previous common definition as growth in a student test score over time, Connecticut's new definition of student growth allows room for non-test-related measures.

Even states that explicitly impose or encourage statewide use of VAMs most often offer such models as off-the-shelf options for local districts lacking the resources or interest to develop their own models. States like Maine, for example, certainly encourage a VAM (again, see Figure 1), but they offer two growth models from which local districts can choose. Both Maine models consider student growth an important aspect for measuring educator effectiveness, but one model uses a VAM while the other employs student learning objectives (SLOs) as measurement tools. SLOs were common before ESSA, but usually as part of a teacher evaluation system that included VAM output as well. In Maine's case, SLOs are effectively replacing VAMs. The trend here is states allowing local control, but also providing ready-made models for districts that do not have resources to create their own models.

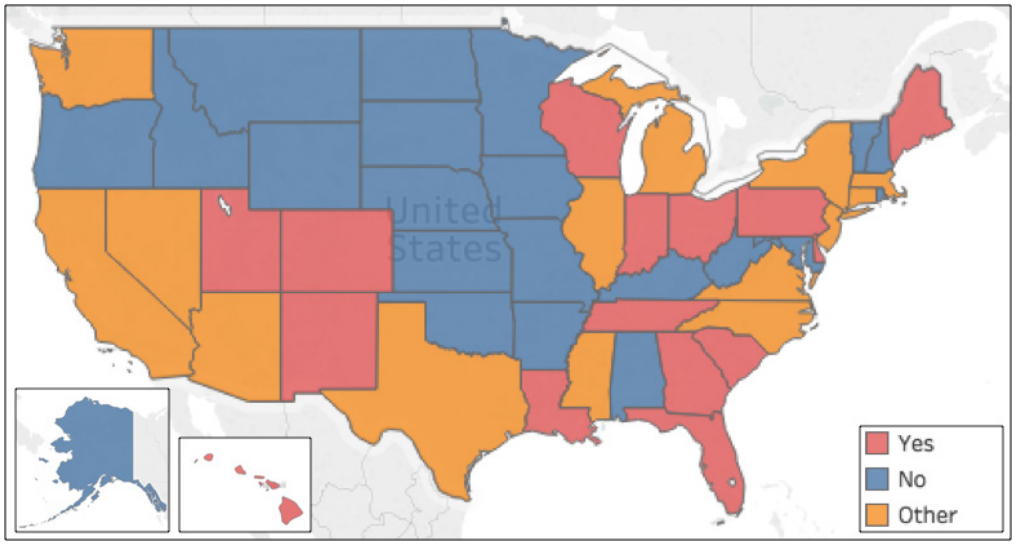


Figure 1. States that use growth models or VAMs as part of their teacher evaluation systems. "Yes" indicates that the state uses or encourages statewide use of growth models or VAMs

(SLOs are not included in this definition of student growth model). "No" indicates the opposite. "Other" answers included some states which use student tests to drive professional development, but not as a factor in evaluation, as well as many states that allow local control without actively encouraging a growth model or VAM.

Other states have retained their former VAMs, but now intend to use them in novel ways. For example, North Carolina uses and reports scores from the EVAAS, a popular VAM, but the state now intends to use EVAAS results to drive professional development, not as a teacher evaluation measure upon which high-stakes consequences or decisions might rest.

While growth models and VAMs are on the decline, they are still prevalent among local districts and state recommended or endorsed teacher evaluation models.

Trend 2: State-level control means more freedom, but implementation challenges as well.

Because ESSA loosened federal control of teacher evaluation, many states no longer have a one-size-fits-all teacher evaluation system. As a result, many current ESSA plans read like guidelines, not rules, deferring to local districts to make more choices about models, implementation, execution, and the like. The firmness of such guidelines, however, also varies widely. California's system, for example, is on one end of the extreme, offering what seems to be the most liberty and freedom to local education authorities, with the state department of education website also yielding a list of helpful resources. Maine, on the other hand, provides a choice of five approved systems from which to select, with all systems including student learning and growth as part of their summative evaluation for teacher evaluation purposes.

One common concern evidenced was the implementation challenges associated with offering or supporting so many different types of teacher evaluation systems within one state. Policymakers noted concerns that those implementing the systems (for example, conducting teacher observations or interpreting assessment data) would not have adequate training. Hence, no matter what elements any states required or recommended, or what measures more open teacher evaluation system might contain, implementation is a paramount concern.

Trend 3: ESSA plans contain more language about supporting teachers by emphasizing formative teacher feedback and de-emphasizing system-level summative evaluations with high stakes consequences.

The rhetoric surrounding teacher evaluation has changed: language about holding teachers accountable for their value-added effects, or lack thereof, is less evident in post-ESSA plans. Rather, new plans make note of providing data to teachers as a means of inciting improvement, essentially shifting the purpose of the evaluation system away from summative and toward formative usefulness. Specifically, 31 out of 51 states and the District of Columbia

(≈61%) cite formative use of data as a part of their evaluation systems. Hawaii's ESSA state plan, for example, states that "formative instructional practices and data teams to foster collaboration among teachers" are priority areas.

Additionally, via the survey, 12 of the 36 (≈33%) state department personnel who answered an open question about the strengths of their teacher evaluation system mentioned formative feedback as a strength. Consequently, unless explicitly stated otherwise, the language of states' post-ESSA plans implies that the teacher evaluation systems are meant not to justify teacher tenure or termination decisions but to provide feedback to teachers. The use of multiple measures (such as observation data, survey data, test score data, and self-assessment data) seems a means of ultimately helping teachers improve their pedagogy and practice.

V. Recommendations

This brief should not be read as a specific direction for creating any assessment or teacher evaluation system, but rather as a showcase of states' new (or not-so-new) ideas, post-ESSA—a state-of-the-states report.

Though each state is unique, and each state faces unique issues related to state-level assessments and teacher evaluation systems, researchers' analyses of the ESSA plans of all 50 states and the District of Columbia reveal some interesting and noteworthy trends. These, in turn, lead to the recommendations below.

Recommendations

Based on findings from the analyses and survey information, it is recommended that state policymakers:

1. Take advantage of decreased federal control by formulating revised assessment policies informed by the viewpoints of as many stakeholders as feasible. Such informed revision can help remedy earlier weaknesses, promote effective implementation, stress correct interpretation, and yield formative information.

2. Ensure that teacher evaluation systems rely on a balanced system of multiple measures, without disproportionate weight assigned to any one measure.

3. Emphasize data useful as formative feedback in state systems, so that specific weaknesses in student learning can be identified, targeted and used to inform teachers' professional development.

4. Mandate ongoing research and evaluation of state assessment systems and ensure that adequate resources are provided to support evaluation.

5. Set goals for reducing proficiency gaps and outline procedures for developing strategies to effectively reduce gaps once they have been identified. Finding which interventions work best to reduce proficiency gaps will take some experimentation. Have a procedure in place to develop these interventions.

Notes and Sources

1 No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

2 Race to the Top (RttT) Act of 2011, S. 844--112th Congress. (2011). Retrieved May 16, 2018, from http://www.govtrack.us/congress/bills/112/s844

3 Hill, D.M., & Barth, M. (2004). NCLB and teacher retention: who will turn out the lights? *Education and the Law, 16*(2-3), 173-181.

4 Amrein-Beardsley, A., & Close, K. (under review). Teacher-level value-added models (VAMs) on trial: Empirical and pragmatic issues of concern across five court cases. *American Educational Research Journal.*

5 See for example: Koedel, C., Mihaly, K., & Rockoff, J.E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

6 American Educational Research Association (AERA) Council (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, X*(Y),1-5. doi:10.3102/0013189X15618385 Retrieved May 16, 2018, from http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html

7 American Statistical Association (ASA) (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA. Retrieved May 16, 2018, from http://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

8 Haertel, E.H. (2013). *Reliability and validity of inferences about teachers based on student test scores* (14th William H. Angoff Memorial Lecture). Princeton, NJ: Educational Testing Service (ETS);

Hill, H.C., Kapitula, L. & Umlan, K. (2011, June). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831. https://doi.org/10.3102/0002831210387916;

Kane, M.T. (2017). *Measurement error and bias in value-added models*. Princeton, NJ: Educational Testing Service (ETS) Research Report Series. doi:10.1002/ets2.12153 Retrieved May 16, 2018, from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12153/full

9 Wong, K.K. (2015). Federal ESEA waivers as reform leverage: Politics and variation in state implementation. *Publius: The Journal of Federalism*, *45*(3), 405-426.

10 Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

11 US Department of Education. (2012). *Elementary and Secondary Education Act (ESEA) flexibility*. Washington, DC: Retrieved May 16, 2018, from https://www.ed.gov/esea/flexibility

12 Braun, V, & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 93. https://doi.org/10.1191/1478088706qp063oa

13 See https://bellwethereducation.org/sites/default/files/Bellwether_ESSAReview_ExecSumm_1217_Final.pdf and https://edtrust.org/resource/trends-state-essa-plans/

14 No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

15 Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: The University of Chicago Press;

Nichols, S. L. & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press;

Grodsky, E.S., Warren, J.R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971-2004. *Educational Policy, 23*, 589-614. https://doi.org/10.1177/0895904808320678

16 Hanushek, E.A., & Raymond, M.E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297–327. https://doi.org/10.1002/pam.20091;

Stotsky, S., Bradley, R., & Warren, E. (2005). School-related influences on grade 8 mathematics performance in Massachusetts. *Third Education Group Review, 1*(1),1-32;

Winters, M.A., Trivitt, J.R., & Greene, J.P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29*, 138-146.

17 Barnett, J.H., & Amrein-Beardsley, A. (2011). Actions over credentials: Moving from highly qualified to measurably effective [Commentary]. *Teachers College Record*. Retrieved May 16, 2018, from http://www.tcrecord.org/Content.asp?ContentID=16517

18 The main differences between growth models and value-added models (VAMs) are how precisely estimates are made and whether control variables are included. Different than the typical VAM, for example, the student growth percentiles (SGP) model is more simply intended to measure the growth of similarly matched students to make relativistic comparisons about student growth over time, without any additional statistical controls (e.g., for student background variables). Students are, rather, directly and deliberately measured against or in reference to the growth levels of their peers, which de facto controls for these other variables. Thereafter, determinations are made in terms of whether students increase, maintain, or decrease in growth percentile rankings as compared to their academically similar peers. Accordingly, researchers refer to both models as generalized VAMs throughout the rest of this manuscript unless distinctions between growth models and VAMs are needed or required.

19 Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record*, *116*(1), 1-32;

Betebenner, D.W. (2011, April). *Student Growth Percentiles*. National Council on Measurement in Education (NCME) Training Session presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

20 Betebenner, D.W. (2011, April). *Student Growth Percentiles*. National Council on Measurement in Education (NCME) Training Session presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

21 Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record*, *116*(1), 1-32.

22 US Department of Education. (2012). *Elementary and Secondary Education Act (ESEA) flexibility*. Washington, DC: Retrieved May 16, 2018, from https://www.ed.gov/esea/flexibility

23 Harris, D.N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press;

Ho, A.D., Lewis, D.M., & MacGregor Farris, J.L. (2009, December). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(4), 15-26.

24 Education Week (2015, see full citation below) illustrated that there were 14 (although there were actually 15) lawsuits filed, in progress, or completed across the nation at the time this article was published. These 15 cases are/were located across seven states: Florida (n=2), Louisiana (n=1), Nevada (n=1), New Mexico (n=4), New York (n=3), Tennessee (n=3), and Texas (n=1), with plaintiffs of all of these cases listing the high-stakes

consequences attached to teachers' value-added indicators of principal concern (e.g., merit-pay in Florida, Louisiana, and Tennessee; tenure in Louisiana; termination in Houston, Texas, and Nevada; and other "unfair penalties" in New York). See Education Week (2015). *Teacher evaluation heads to the courts*. Retrieved May 16, 2018, from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

25 Moore Johnson, S. (2015). Will VAMS reinforce the walls of the egg-crate school? *Educational Researcher, 44*(2), 117-126. https://doi.org/10.3102/0013189X15573351

26 See also Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge;

Baker, B.D., Oluwole, J.O., & Green, P.C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives, 21*(5), 1-71. Retrieved May 16, 2018, from http://epaa.asu.edu/ojs/article/view/1298;

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010, August 27). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute (EPI). Retrieved May 16, 2018, from www.epi.org/publications/entry/bp278;

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS Education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives*. Retrieved May 16, 2018, from http://epaa.asu.edu/ojs/article/view/1594;

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? *Educational Researcher, 44*(2), 132-137. https://doi.org/10.3102/0013189X15575346;

Hill, H.C., Kapitula, L, & Umland, K. (2011, June). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. https://doi.org/10.3102/0002831210387916;

Kappler Hewitt, K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives, 23*(76), 1-49. Retrieved May 16, 2018, from http://epaa.asu.edu/ojs/article/view/1968;

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

27 Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher demographics and evaluation: A descriptive study in a large urban district*. Washington DC: U.S. Department of Education. Retrieved May 16, 2018, from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2017189.pdf;

Steinberg, M.P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293-317. https://doi.org/10.3102/0162373715616249 Retrieved May 16, 2018, from http://static.politico.com/58/5f/f14b2b144846a9b3365b8f2b0897/study-of-classroom-observations-of-teachers.pdf;

Whitehurst, G.J., Chingos, M.M., & Lindquist, K.M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brookings Institution. Retrieved May 16, 2018, from https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-Classroom-Observations.pdf

28 Uttl, B., White, C.A., & Gonzalez, D.W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22-42.

29 Long, C. (2016). Six ways ESSA will improve assessments. *NeaToday*. Retrieved February 14, 2018, from http://neatoday.org/2016/03/10/essa-assessments/

30 Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record*, *116*(1), 1-32.