SPECIAL ISSUE
Policies and Practices of Promise
in Teacher Evaluation

education policy analysis
archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 28 Number 58 April 13, 2020 ISSN 1068-2341

Putting Teacher Evaluation Systems on the Map: An Overview of State's Teacher Evaluation Systems Post–Every Student Succeeds Act

Kevin Close
Audrey Amrein-Beardsley
&
Clarin Collins
Arizona State University
United States

Citation: Close, K., Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of state's teacher evaluation systems post–Every Student Succeeds Act. *Education Policy Analysis Archives, 28*(58). https://doi.org/10.14507/epaa.28.5252 This article is part of the special issue, *Policies and Practices of Promise in Teacher Evaluation*, guest edited by Audrey Amrein-Beardsley.

Abstract: The Every Students Succeeds Act (ESSA) loosened the federal policy grip over states' teacher accountability systems. We present information, collected via surveys sent to state department of education personnel, about all states' teacher evaluation systems post–ESSA, while also highlighting differences before and after ESSA. We found that states have decreased their use of growth or value-added models (VAMs) within their teacher evaluation systems. In addition, many states are offering more alternatives for measuring the relationships between student achievement and teacher effectiveness

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/8/2020
Revisions received: 2/4/2020
Accepted: 2/4/2020

besides using test score growth. State teacher evaluation plans also contain more language supporting formative teacher feedback. States are also allowing districts to develop and implement more unique teacher evaluation systems, while acknowledging challenges with states' being able to support varied systems, as well as incomparable data across schools and districts in effect.

Keywords: education policy; accountability; teacher evaluation

Mapeo de sistemas de evaluación de maestros: Una descripción general de los sistemas de evaluación de maestros del estado después de la Every Student Succeeds Act

Resumen: La Every Student Succeeds Act (ESSA) aflojó el control de la política federal sobre los sistemas de accountability docente de los estados. Presentamos información, recopilada mediante encuestas enviadas al personal del departamento de educación del estado, sobre los sistemas de evaluación docente de todos los estados después de ESSA, al tiempo que destacamos las diferencias antes y después de ESSA. Descubrimos que los estados han disminuido su uso de modelos de crecimiento o de valor agregado (VAM) dentro de sus sistemas de evaluación docente. Además, muchos estados están ofreciendo más alternativas para medir las relaciones entre el rendimiento de los estudiantes y la efectividad de los maestros, además de usar el aumento en el puntaje de la prueba. Los planes estatales de evaluación de maestros también contienen más comentarios de apoyo formativos del maestro. Los estados también están permitiendo que los distritos desarrollen e implementen sistemas de evaluación de maestros más únicos, al tiempo que reconocen los desafíos con que los estados puedan apoyar sistemas variados, así como datos incomparables en las escuelas y distritos vigentes.

Palabras clave: política educativa; accountability; evaluación docente

Mapeamento de sistemas de avaliação de professores: uma visão geral dos sistemas estaduais de avaliação de professores após a Every Student Succeeds Act

Resumo: A Every Student Succeeds Act (ESSA) afrouxou o controle de políticas federais sobre os sistemas estaduais de accountability de professores. Apresentamos informações, compiladas por meio de pesquisas enviadas ao departamento de educação do estado, sobre sistemas de avaliação de professores em todos os estados após a ESSA, enquanto destacamos as diferenças antes e depois da ESSA. Descobrimos que os estados diminuíram o uso de modelos de crescimento ou de valor agregado (VAMs) em seus sistemas de avaliação de professores. Além disso, muitos estados estão oferecendo mais alternativas para medir as relações entre desempenho dos alunos e eficácia dos professores, além de usar a pontuação aumentada no teste. Os planos estaduais de avaliação de professores também contêm um feedback mais formativo e favorável aos professores. Os Estados também estão permitindo que os distritos desenvolvam e implementem sistemas de avaliação de professores mais exclusivos, reconhecendo os desafios de que os estados podem apoiar sistemas variados, além de dados inigualáveis nas escolas e distritos existentes.

Palavras-chave: política educacional; accountability; avaliação do professor

The Policy Topography

Six years before the publication of this article, Collins and Amrein-Beardsley (2014) researched and presented an overview of states' teacher evaluation systems throughout the US after the passage of Race to the Top, a program used to incentivize states into reforming their teacher evaluation systems, primarily via states' consequential uses of data that linked teacher performance to their students' test scores (2011, with data collected in 2012). This descriptive study is an update in the wake of the federal government passing the Every Student Succeeds Act (ESSA, 2016) which eliminated much of the federal role in enforcing test-based accountability across states' teacher evaluation systems. As stated in Ross and Walsh's recent NCTQ report (2019):

> The U.S. Congress's reauthorization of the Elementary and Secondary Education Act of 1965 (ESEA) as the Every Student Succeeds Act (ESSA) in 2015 marks a notable inflection point. ESSA's enactment signaled the end of a period of heightened federal activity that included two initiatives, Race to the Top and ESEA flexibility, both of which incentivized states to develop and implement more objective teacher and principal evaluation systems. (p. 3).

ESSA indicated that states would have more freedom to alter their teacher evaluation policies while (re)embracing more local control (Klein, 2016). However, the rhetoric surrounding ESSA may now be at odds with the current course of teacher evaluation development in which states have already invested significant financial and human resources developing teacher evaluation systems based on previous federal incentives (Jones, Khalil, & Dixon, 2017). In others words, despite the intention or ESSA, some states may be staying the prior course despite the passage of ESSA for a multitude of reasons which may be as varied as the states themselves (Slotnik, Bugler, & Liang, 2016).

The specifics of ESSA gave states more freedom to interpret federally mandated concepts, such as including quantitative or test-based "data on student growth…as a significant factor" of their teacher evaluation systems (e.g., using growth or value-added models, henceforth referred to as "VAMs"; USDOE, 2012). ESSA also allowed states and districts to develop homegrown teacher evaluation systems that used alternative methods and measures to evaluate and attribute student growth to teachers and their effects. However, it is unclear whether states are, in practice, reducing the use of VAMs in teacher evaluation systems or continuing to use VAM output, combined with other measures, in consequential ways. It is also unclear whether states are using the new-found flexibility provided by ESSA to ameliorate what many have argued are the harmful side effects of VAM use (see, for example, *Education Week,* 2015) and the harmful effects of educational accountability that also characterized NCLB.

A study by the National Council on Teacher Quality (NCTQ), for example, indicated that states are not making huge changes post–ESSA (Walsh, Joseph, Lakis, & Lubell, 2017). Researchers in that study used handbooks, guidelines, state websites, and references to legislation to assess such changes. For this study, we collected survey and interview data to investigate how and to what extent states have changed the purposes of, as well as their actual teacher evaluations systems pre– and post–ESSA; the degree to which states are, in practice, reducing the use of VAMs in teacher evaluation systems; and the degree to which states are actually using VAMs in consequential ways.

Re-Surveying the Terrain

The purpose of this article, accordingly, is to provide an updated overview of all states' teacher evaluation systems following the passage of ESSA (2016), and to also include insights into how state department of education personnel view the strengths and weaknesses of their new and re-reformed teacher evaluation systems. Our two-fold objectives for this study draw strength from providing both an outside view (i.e., a summary of state plans post–ESSA) and an inside view (i.e., an aggregated analysis of common perceptions from the personnel who created and oversee states' evaluation systems).

We collected the same general data as in Collins and Amrein-Beardsley's (2014) prior study, but we asked refined questions to better match the current context. For example, in the earlier study (Collins & Amrein-Beardsley, 2014), VAMs and the high-stakes consequences tied to teacher evaluation systems that relied heavily on VAM output dominated the discourse around states' teacher evaluation systems. However, because ESSA allowed states more leniency over their states' teacher evaluation systems, researchers sought more holistic information in this study about states' teacher evaluation measures, including but not limited to only VAMs.

We present findings in visual (e.g., a series of maps) and raw versions (e.g., a table displaying data on each state's current teacher evaluation measures) so readers can directly access states' information. Comparable data before and after ESSA is also presented as a series illustrating changes over time including a table detailing how certain features of each state's teacher evaluation systems have changed post–ESSA. Prior to presenting findings, though, it is important to review the relevant literature used to both situate and frame this study.

Relevant Literature

With the passage of the NCLB (2002), the early 2000s throughout the US marked a new era in educational accountability policies, with federal policies increasingly promoting accountability-based systems that held students, teachers, and schools responsible for improved student achievement results. Some research indicated that teachers affected student performance and that teacher performance differed within schools (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Despite this, most teacher evaluation systems, as based primarily on principal observation, indicated that almost all teachers received satisfactory results (Weisberg, Sexton, Mulhern, & Keeling, 2009). Hence, the theory of change was that by holding schools, teachers, and students accountable for meeting higher standards, as measured by student performance on standardized assessments, administrators would supervise public schools better, teachers would teach better, and students would take their learning more seriously. As a result, students would learn and achieve, or rather progress more, particularly in the lowest performing schools.

However, many researchers now agree that NCLB did not meet its intended effects (100% student mastery of higher standards by 2014). More specifically, research suggests that since the passage of NCLB, many students, especially those in the country's lowest performing schools, have been increasingly susceptible to unprofessional test-based practices including teaching to the tests (not to be confused with teaching to the standards); teaching using scripted and prefabricated curricula to ensure that what is taught aligns with what is tested; teaching test preparation, test practice, and test rehearsals instead of curricular content; teaching while hyper-emphasizing the rote memorization of facts and basic skills likely to be on tests; narrowing the curriculum to match the content and concept areas tested; and, related, teaching the tested subject areas that "count" (i.e., mathematics and reading/language arts) while marginalizing or even eliminating other curricular areas and activities that do not "count" on high-stakes tests (i.e., social studies, sciences, art, music, physical education, library sciences, and recess; see, for example, Amrein & Berliner, 2002; Haney, 2002; Nichols & Berliner, 2007).

Also, typically low-scoring students, including inordinate numbers of non-English proficient and special education students have been purged (i.e., expelled, suspended, or simply excused) from school during test administrations to keep them from participating and pulling test scores down. Students have also been counseled out of school, convinced to explore other options (e.g., alternative, "last chance," or adult education schools), or persuaded to strive for General Education Diplomas (GEDs) instead of traditional high school certificates. Eliminating undesirable students eliminates their scores; the scores that if included or preserved would pull composite test scores down (see, for example, Amrein-Beardsley & Berliner, 2002; Haney, 2000; Nichols & Berliner, 2007).

Students whom educators have deemed the least likely to post high enough test scores, the same students as mentioned above, have also been academically shunned. This has occurred particularly during the weeks leading up to high-stakes tests as students are often perceived by educators being held accountable as the most hopeless, and hence, the most undesirable when test scores punitively matter. Undesirable students have also been known to be retained in grade or credit hours to keep them from being eligible for high stakes testing cycles (e.g., by thwarting progression in high school whereas sophomores/juniors might not be eligible to test in their sophomore/junior year; see for example Haney, 2000). In some cases, undesirable students have altogether disappeared from school rosters when administrators have created rosters and registered students for high stakes testing purposes (see also Amrein & Berliner, 2002; Nichols & Berliner, 2007.

Otherwise, underperforming students have been wrongly moved into exempt categories (e.g., special education and English Language Learner [ELL] categories), as misclassifying these students will prevent them from dragging down the performance of the teachers or the schools as a whole (Amrein & Berliner, 2002; Haney, 2000). Recognizing this as an issue, the federal government started mandating minimum rates of test participation (NCLB, 2002), but it seems such practices are still occurring.

Conversely, educators have focused inordinately on the students who are on the edge of passing high-stakes tests. The belief here is that if educators teach to the test well enough, these students just might clear the cut scores and pass, thus helping to bump composite test scores, even if ever so slightly upwards. Educators have used "selective seating" practices in which the students expected to post high scores are seated among the students expected to post low scores, covertly encouraging cheating. Educators have also overtly cheated, for example, by erasing and changing students' incorrect answers to correct, explicitly giving students correct answers, persuading students to revisit incorrect answers, and the like. Such cheating instances have been widely publicized, for example, in Atlanta and Washington D.C. (Perry & Vogell, 2009; Rhee, 2011) as well as in the Arizona (Amrein-Beardsley, Berliner, & Rideau, 2010; see also Toppo, Amos, Gillum, & Upton, 2011).

Likewise, some argue that these unintended effects (as well as others; see also Darling-Hammond, 2007; Figlio & Getzler, 2006) may have outweighed some of the positive effects noted, including but not limited to an increased focus on measuring and monitoring the gaps between marginalized and non-marginalized student populations (see, for example, Grodsky, Warren, & Kalogrides, 2009; Koretz, 2017; Nichols & Berliner, 2007). The results, of course, are controversial with others arguing that the NCLB era positive effects outweighed the negative effects (Dee & Wyckoff, 2015; Winters, Trivitt, & Greene, 2010; see also Hanushek & Raymond, 2005; Stotsky, Bradley, & Warren, 2005).

Regardless, after collectively acknowledging some of the issues with NCLB, the federal government used federal funds again to entice states and districts to move in new directions. Consequently, the federal government (e.g., via Race to the Top, 2011 and the aforementioned

NCLB waivers [USDOE, 2014][1]) incentivized states to adopt new and improved tests (e.g., those developed by the Partnership for Assessment of Readiness for College and Careers [PARCC] or Smarter Balanced Assessment Consortium [SBAC]), to adopt and implement new and improved educational policies, and to use both (i.e., improved tests and improved test-based accountability policies) to hold teachers accountable for their students' growth in learning and achievement over time. 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 students' growth in learning over time.

Soon after Race to the Top (2011) was underway, 40 states and the District of Columbia were using, piloting, or developing some type of VAM, again, as federally incentivized (Collins & Amrein-Beardsley, 2014). The tests required under NCLB were used across states for measuring teacher-level value-added. The most common open-source VAM was the student growth percentiles (SGP) model (Betebenner, 2009, 2011), with multiple states adopting or endorsing it for teachers statewide (i.e., Arizona, Colorado, Georgia, Massachusetts, and Washington). The SGP model compares the growth from one year to the next with similar peers.[2] The most common proprietary model was the Education Value-Added Assessment System (EVAAS; Sanders & Horn, 1994; Sanders, Wright, Rivers, & Leandro, 2009; SAS Institute Inc., n.d.), with five states adopting it statewide (i.e., North Carolina, Ohio, Pennsylvania, South Carolina, and Tennessee). Unlike the SGP model, the EVAAS model is a proprietary statistical model with an unknown algorithm for measuring the impact of teachers on student learning. The most common high-stakes consequences being attached to systems that included VAM results included but were not limited to teacher tenure, termination, and teacher compensation or merit pay (Collins & Amrein-Beardsley, 2014).

While all teacher evaluation systems adopted and implemented at this time included at least one other indicator or measure of teacher effectiveness (i.e., systemic classroom observations of teachers), the primary focus across states was on the objective, assessment-based (and often VAM-based) components to "meaningfully differentiate [teacher] performance…including as a significant factor, data on student growth [in achievement over time] for all students" (USDOE, 2012). Some research supported the use of such teacher evaluation systems (Chetty, Friedman, & Rockoff, 2014a; Kane & Staiger, 2012). This strategy was written into federal policy and subsequently implemented across the nation, although some states (e.g., Florida, Louisiana, Nevada, New Mexico, New York, Tennessee, and Texas) valued or systemically weighted student growth (i.e., teachers' value-added) much more heavily in their systems than others (e.g., California, Connecticut, Vermont, Washington, and Wisconsin).

---

[1] It is important to also note here that the federal government also granted states waivers from *not* meeting No Child Left Behind (NCLB, 2002) goals for their students to reach 100% academic proficiency by 2014 *if* states also created and adopted stricter teacher evaluation systems as based, at least in part, on VAMs (US Department of Education, 2014). Most states applied for these waivers, also making shifting most state's teacher evaluation systems to their highest-accountability versions.

[2] 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.

Around this time, research on VAMs, especially in conjunction with teacher evaluation systems, increased heavily. VAMs, in the simplest of terms, classify teachers' effectiveness according to teachers' statistically measurable (and purportedly) causal impacts on their students' standardized test scores over time. While there is debate about the extent to which VAMs can be used to separate out a teacher's impact from other classroom-level factors (see, for example, Rothstein, 2009, 2010), the intent of VAMs is to help to identify teachers whose students outperform their projected levels of growth as effective or of "value-added" and teachers whose students fall short as the inverse (Sanders, 2003). Views on such assessment-based systems are controversial when attaching high-stakes consequences to such measures of teacher effectiveness (American Statistical Association [ASA], 2014; American Educational Research Association [AERA] Council, 2015; Baker et al., 2010; see also Harris, 2011; Ho, Lewis, & MacGregor Farris, 2009).

These controversial views led to court challenges to states' VAM-based teacher evaluation systems (i.e., in Florida, Louisiana, Nevada, New Mexico, New York, Tennessee, and Texas; see *Education Week,* 2015).[3] Plaintiffs argued the following main points of criticism regarding VAM models within teacher evaluations systems including that VAMs can be: (1) unreliable, whereby current research suggests that teachers classified as "effective" one year will have a 25%-59% chance of being classified as "ineffective" the next year, or vice versa, with other permutations possible (Chiang, McCullough, Lipscomb, & Gill, 2016; Martinez, Schweig, & Goldschmidt, 2016; Schochet & Chiang, 2013; Shaw & Bovaird, 2011; Yeh, 2013); (2) invalid, whereby very limited research evidence supports the claim that VAMs can be used to draw accurate inferences about the extents to which different teachers *cause* changes (i.e., add value) in a collective groups of students' test performance over time (see, for example, Amrein-Beardsley, 2008; Braun, 2005, 2015; Hill, Kapitula, & Umland, 2011); (3) biased, whereby current research suggests that, almost regardless of the sophistication of the statistical controls used to block bias, VAM-based estimates sometimes present biased results, especially when relatively homogeneous sets of students (i.e., ELLs, gifted and special education students, free-or-reduced lunch eligible students) are non-randomly concentrated in schools and teachers' classrooms (Baker et al. 2010; Capitol Hill Briefing, 2011; Collins, 2014; Green, Baker, & Oluwole, 2012; Kappler Hewitt, 2015; Koedel, Mihaly, & Rockoff, 2015; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Rothstein & Mathis, 2013); (4) not transparent, with the main issue being that VAM-based estimates do not often make sense to those at the receiving ends of the estimates (e.g., teachers and principals) and, subsequently, these same groups are reportedly quite-to-very unlikely to use VAM-based output for formative purposes (see, for example, Eckert & Dabrowski, 2010; Gabriel & Lester, 2013; Goldring et al., 2015; Graue, Delaney, & Karch, 2013); and (5) unfair, with the fundamental issue being that states and districts can only produce VAM-based estimates for approximately 30-40% of all teachers, leaving the other 60-70% (which sometimes includes entire campuses of teachers) ineligible under comparable evaluation and accountability systems (Baker, Oluwole, & Green, 2013; Gabriel & Lester, 2013; Harris, 2011).

In light of the recent critical research and court cases regarding VAMs, observational systems used for similar teacher evaluation purposes, which were also deeply criticized and subsequently spurred some of the federal government's reforms (Weisberg et al., 2009; see also Kraft & Gilmour, 2017), are now even more common across states' new and re-reformed (i.e., post–ESSA, 2016) teacher evaluation systems (Ross & Walsh, 2019; Steinberg & Donaldson, 2019). They are still, however, also confronting their own sets of empirical issues. Such issues include but are not limited to whether the observational systems are psychometrically sound for such purposes, how output from observational systems might be biased by the supervisors observing teachers in practice, and how output might also be biased by contextual factors like the types of students with whom a teacher works, how a teacher's gender interplays with his/her students' gender, and other factors (Bailey, Bocala, Shakman, & Zweig, 2016; Geiger & Amrein-Beardsley, 2017; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014). The same sorts of potential biases seem to hold true with student surveys, regardless of whether also used to evaluate teachers in Pre-K or evaluate instructors in higher education, given selection biases.[4]

---

[3] *Education Week* (2015) 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).

Nonetheless, the new freedom that ESSA (2016) has afforded states means they could be (and anecdotally are) moving away from such high-stakes and assessment-based accountability models, especially from those based primarily on VAMs. Ideal components of a teacher evaluation system would include standards-based teacher observations across the year, systems that provide timely formative feedback, multiple sources of evidence of student learning, and greater collaboration between teachers or between teacher and administrators (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). Essentially, ideal components of a teacher evaluation system would reflect the latest standards of educational and psychological testing, meaning the results would reliable, valid, fair, unbiased, and transparent (AERA, NCME, APA, 2014). However, policymakers need also be wary of the unintended consequences caused by imposing new measures. The potential for unintended consequences is one reason that Darling-Hammond et al. (2012) recommend teacher evaluation systems that encourage greater collaboration between teachers or between teacher and administrators.

Accordingly, this study aims to uncover whether states are actually taking advantage of the purported flexibility within ESSA (2016) policy and to what extent, for example, by uncovering whether states are moving in new directions, away from such common-because-they were-federally-incentivized models, and away from using VAMs as their primary teacher evaluation and accountability measures.

Methodology

We conducted a survey research study using an electronic survey along with phone interviews to contact non-respondents, to follow-up for clarification, and for validation purposes. We engaged these methods to gather central and supplementary information about all states' restructured teacher evaluation systems post–ESSA. We collected all survey- and phone-based information from state department of education personnel directly. Some state department personnel referred us to pertinent state-level documents (e.g., state policies and other legislative pieces, as well as state ESSA plans) online. Additionally, for states that did not respond to survey invitations or phone calls we evaluated ESSA plans and referred to state education department websites.

The four research questions that we examined for this study were: (1) What measures are being used by each state to evaluate teachers? (2) How have states' teacher evaluation systems changed following the adoption of ESSA? (3) What do state personnel see as the strengths and weaknesses of their post–ESSA teacher evaluation systems? (4) How have state personnel's perceptions of the strengths and weaknesses changed post–ESSA?

---

[4] Response bias is of concern when the sample of responses obtained is not representative of the population intended to be analyzed or intended to be represented by the sample of responses obtained. See also Uttl, White & Gonzalez,2017).

Participants

Study participants included state education personnel from every state and the District of Columbia, hereafter generally referred to in plural as states (*n* = 51), representing those most knowledgeable of each state's teacher evaluation system post–ESSA. To locate the most knowledgeable personnel to participate in this study, we first searched for state personnel online looking at job titles relating to teacher evaluation, teacher quality, or teacher accountability. We then emailed or called to verify they were the best source of information for our study or if we should contact a different source. In some cases, where we did not find appropriate job titles, we simply called the state department of education and asked with whom we should contact. Contacts were provided a description of our study along with a description of the survey before choosing to do the study to ensure that we ultimately communicated with those who were the most knowledgeable.

Participants ultimately included leaders and directors of states' teacher quality departments, leadership divisions, evaluation offices, and accountability and assessment divisions. Of the 51 departments contacted, personnel representing 34 (67%) states responded to the online survey and personnel representing four (8%) states answered survey questions via phone interviews. Additionally, representatives from four (8%) state departments did not answer the questions specifically, so we referred to online resources instead. In sum, personnel from 42 (82%) state departments of education responded via survey, and for the other nine departments (18%), we captured states' missing information by reading publicly available state websites and states' ESSA plans. Accordingly, we indicate their sources of information by state (e.g., whether information was collected through personal contact or through state websites) within the findings presented.

Survey Instrument

We developed the survey instrument used to collect state data over the course of three months in order to increase the validity, accuracy, and relevancy of the instrument, but also to increase the likelihood of states' participation. To develop the survey instrument, we developed overarching questions based on Collins and Amrein-Beardsley's (2014) study prior to ESSA. Thereafter, we developed additional questions given the aforementioned, and expanded goals and objectives for this study.

Following guidelines for effectively conducting survey research studies (Kelley, Clark, Brown, & Sitzia, 2003), we first conducted content analysis with state department of education personnel within our own state and pilot tested the instrument with three other state personnel and teacher evaluation experts to ensure that the content and format of the survey were clear, comprehensive, and relevant given states' realities and expectations post–ESSA. The pilot tests included observing and asking the participants whether each question made sense, whether their responses were indeed the information we were intending to gather via the survey, and overall feedback on wording, length of survey, and other practical questions. For the states that participated via phone interview or for which we analyzed documents (e.g., states' post–ESSA teacher evaluation plans), we manually input data into the same survey instrument to allow for one primary database which kept all data collected constant, consistent, and comparable. Click here for the full survey instrument that we validated and used for these purposes.

Procedures

We distributed the survey instrument to all state personnel online via Qualtrics Survey Software (2019). As explained prior, data collection also consisted of making phone calls to state personnel in order to encourage the participation of non-respondents, and to also ask clarifying questions, to ensure responses were accurately represented, to verify that nothing had changed from previous communications, and to ensure that states' data were accurate and representative of the current and most up-to-date teacher evaluation situations by state. Again, these data collected via phone interviews were inserted into the same survey instrument as if the person on the phone were completing the survey themselves.

Data Analyses

For the survey items that yielded quantitative information, we calculated frequencies and descriptive statistics. For the survey instrument items that yielded qualitative responses (e.g., items that solicited personnel opinions on the strengths and weaknesses of their teacher evaluation systems), we aggregated these data to protect the anonymity of the state responses. Once aggregated, we followed the methods and procedures outlined in Miles and Huberman (1994) using a sourcebook to "[track] out lawful and stable relationships among social phenomena based on the regularities and sequences that link these phenomena" (p. 174) during the processes of data reduction, data display, and drawing conclusions. Lastly, we used Tableau Software (2019) for constructing map visualizations of the descriptive data for ease in interpretation.

It should be noted here, though, that because state plans often change, some state-level information may have changed between data collection and publication. On the flipside, the reported and perceived strengths and weakness of states' systems from participating personnel may indicate the direction of said changes. Regardless, both of these points should be noted so that consumers do not interpret the forthcoming results as fixed.

Results: The National Landscape

The results section maps onto aforementioned research questions, and within each section we present results in three ways: (1) as aggregate tables, (2) as series of maps, and (3) in prose. We chose to present the results in these ways because the purpose of this paper is to present as complete a picture of the state of states' teacher evaluation systems within the constraints of a journal article. We understand that tables containing information from all 50 states and the District of Columbia would be unwieldy, so we designed the presentation of data in such a way that readers might have direct or immediate access to what we deemed to be the most important results (e.g., via the maps and prose forthcoming). However, we also created easily accessible larger, searchable, and sortable tables including results that provide more in-depth and by-state data that we uploaded online within a set of accessible and anonymous spreadsheets.

Research Question 1: States' Teacher Evaluation Measures

In this section, we break down the results of the survey by each teacher evaluation measure now being used by states including (1) VAMs (defined prior), (2) Teacher-Level Observations (used to purposefully examine teachers' teaching practices in context through systematic processes of data collection, analysis, and reflection; Bailey, 2001), (3) Student Surveys (used to systematically obtain students' opinions about different aspects of their teachers' attitudes, instruction, and pedagogical practices; Geiger & Amrein-Beardsley, 2017), and (4) Student Learning Objectives (SLOs; used to measure teachers' students' growth using one or more traditional [e.g., state-wide standardized tests] or non-traditional assessments [e.g., district benchmarks, school-based assessments, teacher and classroom-based measures]; see Lacireno-Paquet, Morgan, & Mello, 2014; USDOE, n.d., p. 1[5]). For each of these measures, we provide a map illustrating which states adopted which of these measures post–ESSA (2016). This section concludes by presenting an anonymized link to a full table indicating each states' teacher evaluation measures.

Value-added models (VAMs). As stated previously, the state of states' continued uses of VAMs post–ESSA (2016) was unknown, given ESSA rolled back some test- and growth-based mandates for all states' teacher evaluation systems. Findings herein indicate that 15 states explicitly use or encourage state-wide use of VAMs (29%, 15/51), many of which offer VAMs as state-supported or endorsed options for districts that do not have the resources (e.g., budget or personnel hours) to develop a homegrown VAM or growth model. Twenty-two states explicitly do not use or encourage state-wide use of VAMs (43%, 22/51), and 14 states (27%, 14/51) report the use of "other" approaches regarding VAMs (See Figure 1). For the roughly one-third of states claiming they now use or endorse "other" approaches, 10 of those states (20%, 10/51) reported that they had passed these choices onto districts in the name of local control (i.e., local educational authorities such as school districts can choose to use VAMs), two states reported that VAMs were now being used formatively or for only informative purposes (4%, 2/51), one state reported that their state's VAM was still in development (2%, 1/51), and one state's current situation in this regard remains unknown (2%, 1/51).

Examples of states that offer, but do not mandate state-wide VAMs include Maine which has two models from which local districts can choose to evaluate teacher performance. One model uses a VAM to measure student growth, and the other uses a SLO as a way to measure student growth. Another state, Texas, emphasizes local control. Their department of education allows student growth to be measured several ways including SLOs, portfolios, district-level pre- and post-tests, and VAMs in state-tested subjects.

Yet other states are still using VAMs, but they are using them in less traditional ways. For example, North Carolina uses and reports scores from the aforementioned EVAAS, but state personnel use the results to drive teacher professional development and no longer as a high-stakes teacher evaluation measure. In fact, in their ESSA plan, North Carolina recommends that student growth scores be discussed with teachers mid-year as a way of checking on progress towards instructional practice goals set at the beginning of the school year. The plan explicitly calls for EVAAS scores to be used to stimulate discussion as one of multiple measures of teacher effectiveness. Put differently, although North Carolina technically encourages the use of one VAM to evaluate its teachers, the state encourages VAMs' formative over summative uses, which was not nearly as prevalent prior to the passage of ESSA (2016; see more on this forthcoming; see also Figure 1).

---

[5] More officially, and according to the USDOE (n.d.), SLOs are flexible objectives that can be set by teachers, administrators, districts, or some combination which evidence student growth. The USDOE states, "It is possible to use large scale standardized tests, even state standards tests for SLOs. However, it is also possible to use other methods for assessing learning, such as end of course exams in secondary courses, student performance demonstrations in electives like art or music, and diagnostic pre- and post-tests in primary grades or other relevant settings" (p. 1).

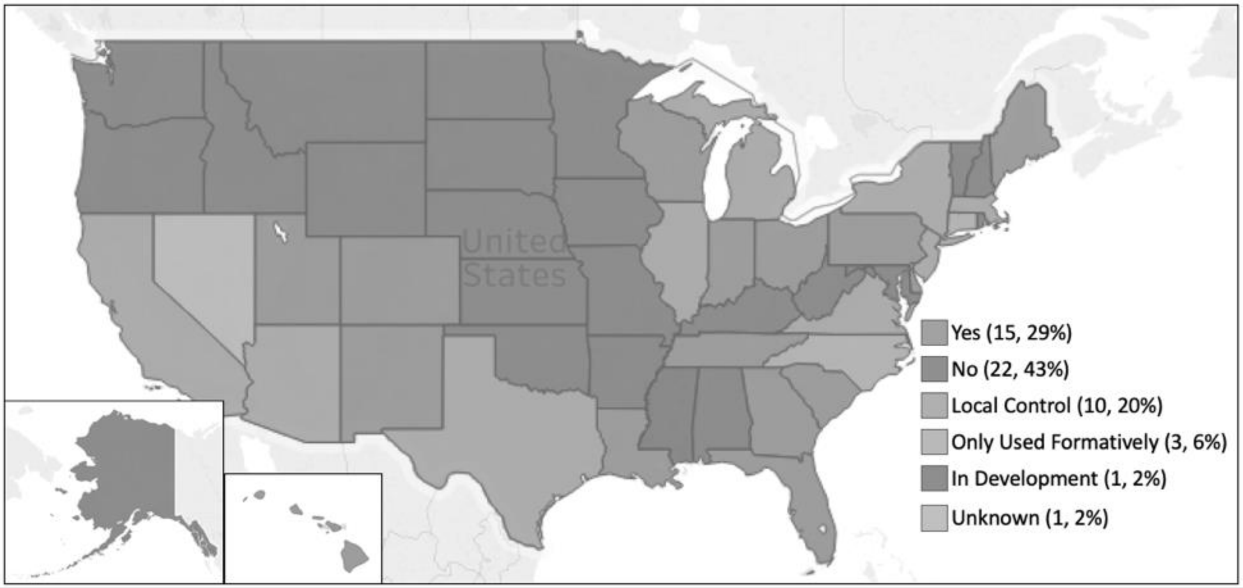

Figure 1. States that use VAMs as part of their teacher evaluation systems (2018).
Note: Fifteen states use VAMs (29%, 15/51), 22 states do not use VAMs (43%, 22/51), 10 states report local control (20%, 10/51), three states use VAMs, but only formatively (6%, 3/51), one state had a VAM that is still in development (2%, 1/51), and one state is unknown (2%, 1/51).

Teacher-level observations. Teacher-level observations are also a dominant feature across states' current teacher evaluation systems with 36 of 51 (71%) states reporting use. States which do not report using teacher level observations, such as Wyoming, along with six other states (12%, 6/51), may ultimately use teacher level observations, given local control to select elements of their teacher evaluation plans; however, they do not explicitly indicate their use, as compared with the 36 states that indicated their widespread use. Additionally, five states (10%, 5/51) explicitly (e.g., via state-level policy) allow for local control in terms of using teacher-level observation systems (see Figure 2).

Of the 36 states in which teacher-level observations are encouraged, 18 of the 36 states (50%) use or encourage the Danielson's Framework for Teaching observational system, or a modified version (Danielson, 2012; Danielson & McGreal, 2000), and 11 of the 36 states (31%) use or encourage the Marzano Causal Teacher Evaluation Model (Marzano & Toth, 2013).[6]

There is some overlap among states that use or encourage Danielson's Framework for Teaching and the Marzano Casual Teacher Evaluation Model, with eight of the 36 states (22%) either using or encouraging both of these models or others. For example, Alabama uses an observation framework based on a combination of its Alabama Quality Teaching Standards and the work of both Danielson and Marzano. Alaska allows local school districts to select from several major frameworks including but not limited to Danielson and Marzano. Other states encourage various observational systems or multiple observation systems including homegrown rubrics that are developed from a state model (8%, 3/36), outside rubrics aligned to a state rubric (8%, 3/36), and the National Institute for Excellence in Teaching's (NIET's) TAP System for Teacher and Student Advancement (11%, 4/36) (NIET, n.d.; see also, Barnett, Rinthapol, & Hudgens, 2014).

---

[6] Briefly, both of these models are based on a specific conceptualization of the elements of teaching. Danielson's Framework for Teaching conceptualizes teaching as a complex activity with four main responsibility domains: a) planning and preparation, b) classroom environments, c) instruction, and d) professional responsibilities. Within each of these domains, the activity of teaching is further broken down into 22 components with 76 subcomponents (Alvarez & Anderson-Ketchmark, 2011). Danielson's observational framework emphasizes collecting evidence based on these componenets, interpreting such evidence, and conducting professional conversations with teachers the evidence (Danielson, 2012). The Marzano Causal Teacher Evaluation Model uses a similar framework based on four domains: a) classroom strategies and domains, b) planning and preparing, c) reflecting on teaching, and d) collegiality and professionalism. Within these domains, like the Danielson Framework, teaching is broken down into 60 elements with the majority falling under the umbrella of classroom strategies and domains (Marzano, 2012).

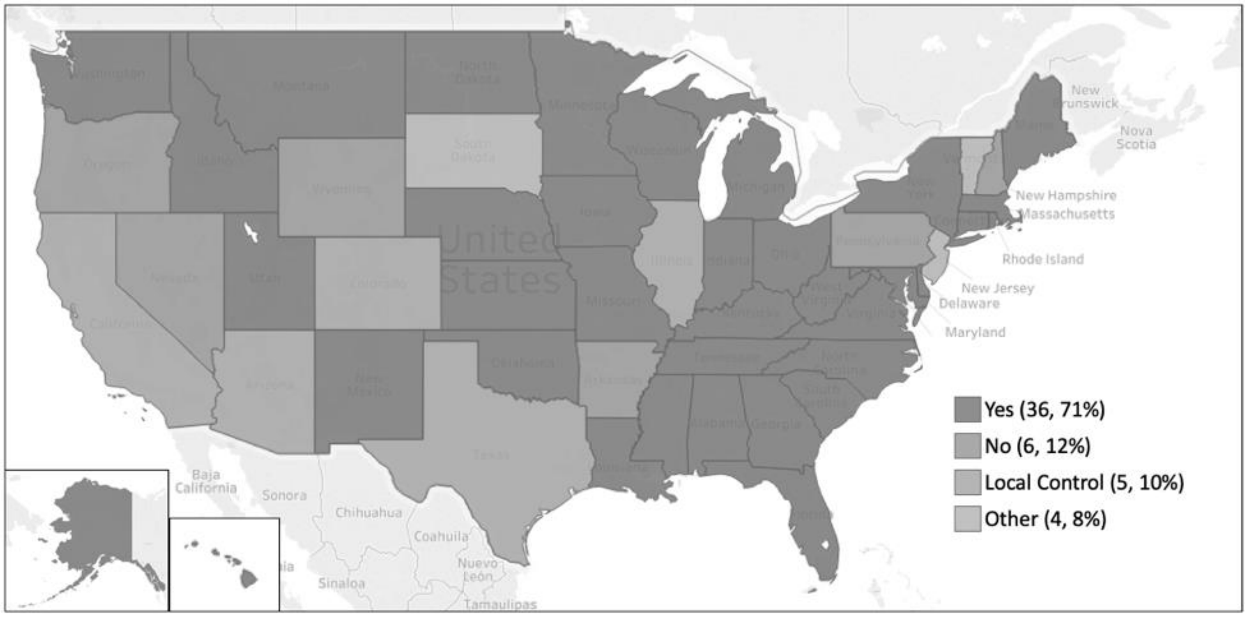

Figure 2. States that include observations as part of their teacher evaluation systems (2018).
Note: Thirty-six states use teacher observations (71%, 36/51), six states do not use teacher observations (12%, 6/51), five states report local control (10%, 5/51), and four are classified as "other" (8%, 4/51).

Student surveys. Student surveys of their teachers are used much less frequently than VAMs and observations, but they are on the rise in terms of development, adoption, and implementation (Geiger & Amrein-Beardsley, in press). Indeed, 14 of 51 states (27%) reported using or encouraging the use of student surveys to evaluate their teachers, and one state (2%), Washington, is currently piloting a state-wide student survey system. While 16 of 51 states (31%) explicitly noted not using or encouraging student surveys, it is evident that teacher evaluation measures are more common now than post–Race to the Top (2011; see also Kane & Staiger, 2012). Additionally, 13 of 51 states (25%) allow local control with regard to student survey systems that can also take many forms. For example, the Colorado Department of Education neither specifies nor recommends specific student surveys; however, their state statute requires that the use of a student survey as a viable option for districts when evaluating their teachers. In other words, local educational authorities can decide whether or not to even use the measure. Arkansas, on the other hand, encourages the use of perceptual data from multiple stakeholders including students, but the formats via which these data are collected are left to local authorities to decide. As not all states clearly distinguish whether they use student surveys, 7 of 51 (14%) states also remain unknown in this regard (see Figure 3).

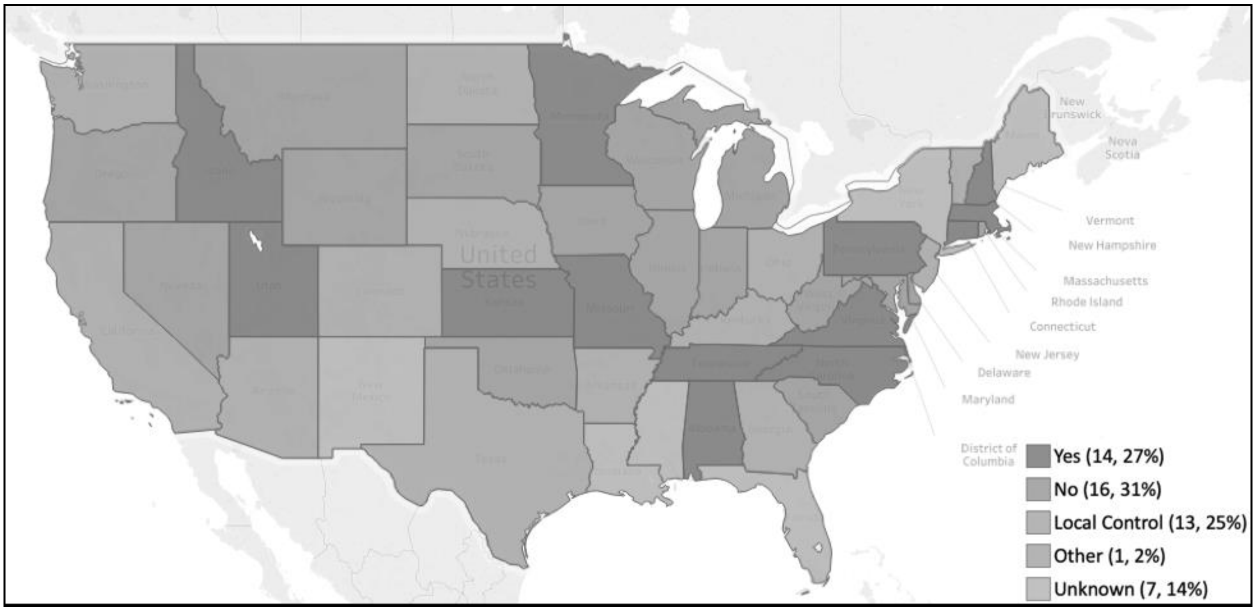

Figure 3. States that include student surveys as part of their teacher evaluation systems (2018).
Note: Fourteen states include student surveys (27%, 14/51), 16 states do not include student surveys (31%, 16/51), 13 states report local control (25%, 13/51), one state is classified as "other" (2%, 1/51), and seven states are unknown in this regard (14%, 7/51).

Student learning objectives (SLOs). More than half of the states (28 of 51 states; 55%) use or encourage SLOs in their teacher evaluation systems with seven of 51 (14%) not explicitly using or encouraging SLOs statewide. Another nine of 51 (18%) use SLOs as a substitute for VAM data for teachers whose subject areas do not align with state tests (e.g., for primary grade and non-core subject area teachers). Three of 51 states (6%) report local control for this indicator including Texas, which encourages teachers to set goals for student learning but does not prescribe that local education agencies use SLOs specifically. Lastly, four of 51 states (8%) do not clearly state whether they use SLOs and are accordingly classified as "unknown" (see Figure 4).

Unlike teacher-level observation frameworks which are relatively well-developed and have been around and in development and refinement for decades, (Sloat, Amrein-Beardsley, & Sabo, 2017), SLOs do not appear to be nearly as well-developed, conventionally used, or established in comparison to all of the other teacher evaluation measures in play across states given these observational frameworks (see also USDOE, n.d.). For example, in Nebraska SLOs are officially encouraged, but their use is not yet widespread. In Nevada, teachers and their supervisors use tools to create Student Learning Goals (SLGs), but the processes by which these are created vary widely by teacher and supervisor. Both practices are akin to what other SLOs might involve or look like, but nowhere are SLGs differentiated from SLOs, even despite their similarities. In Illinois, SLOs are the default teacher evaluation measure. If school districts cannot come to consensus on another so-called growth-based system, 50% of all teachers' overall evaluation scores rely upon their SLO data.

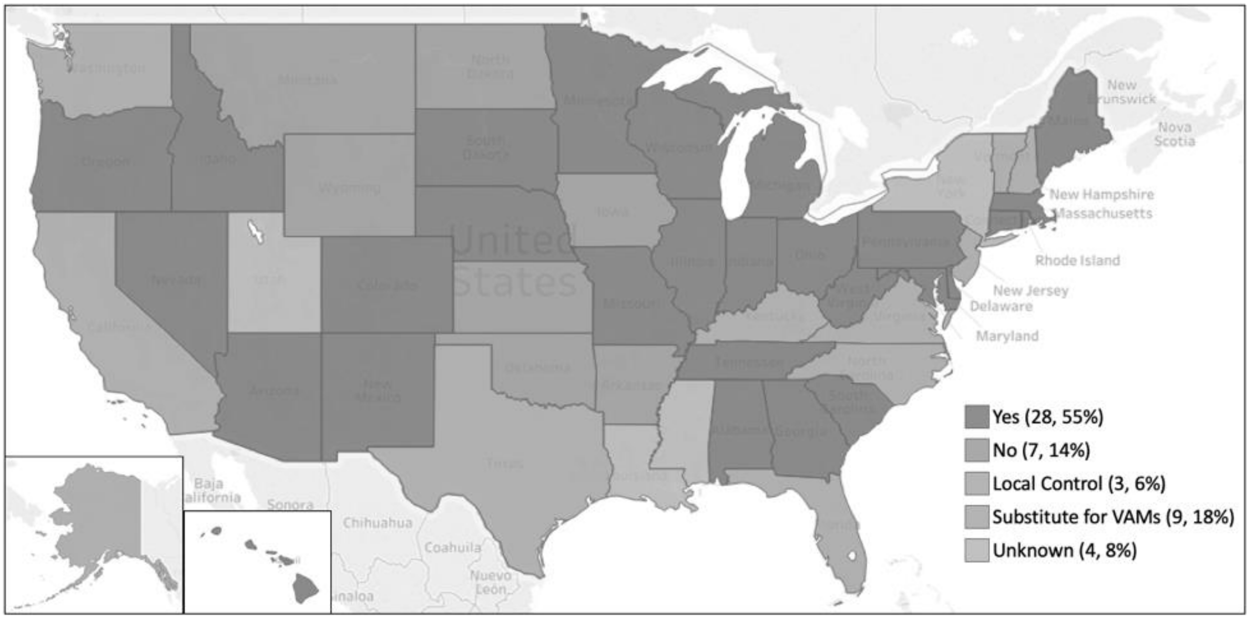
Washington
Montana
North Dakota
Minnesota
Oregon
Idaho
South Dakota
Wisconsin
Wyoming
Iowa
Michigan
Nevada
Utah
Colorado
United States
Illinois
Indiana
Ohio
Pennsylvania
California
Missouri
West Virginia
Kentucky
Virginia
Oklahoma
Arkansas
Tennessee
North Carolina
South Carolina
Arizona
New Mexico
Texas
Alabama
Georgia
Louisiana
Florida
New York
Maine
New Brunswick
Nova Scotia
New Hampshire
Massachusetts
Rhode Island
New Jersey
Delaware
Maryland
Baja California
Sonora
Chihuahua
Coahuila
Nuevo León
Sinaloa
Yes (28, 55%)
No (7, 14%)
Local Control (3, 6%)
Substitute for VAMs (9, 18%)
Unknown (4, 8%)

Figure 4. States that include SLOs as part of their teacher evaluation systems (2018).
Note: Twenty-eight states include SLOs (55%, 28/51), seven states do not include SLOs (14%, 7/51), three states report local control (6%, 3/51), nine states use SLOs as a substitute for VAMs (18%, 9/51), and four states are unknown in this regard (8%, 4/51).

While the preceding section illustrates the results to our first research question in this study, via the use of maps, descriptive statistics, and summary paragraphs, we gathered more detailed information about states' teacher evaluation systems that can be found in Table 1. Again, this online anonymous table includes the state-by-state information that yields much more in-depth information than the figures and text included above. This table includes information collected via the survey instrument such as VAM-specific legislation, types of assessments used to measure student growth, consequences attached to teacher evaluation measures, and percentage of overall teacher evaluation determined by student growth.

Research Question 2: How States' Teacher Evaluation Systems Have Changed Post–ESSA?

The following section transitions from explaining the status of state's current teacher evaluation systems and measures to highlighting how states' systems may have changed since Collins and Amrein-Beardsley (2014) last collected data post–Race to the Top (2011). Again, the 2014 study collected information specifically regarding VAMs and VAM use. Therefore, the information included next includes only comparative data on states' VAM-related information given no other information about states' teacher evaluation measures were collected in Collins and Amrein-Beardsley (2014).

Accordingly, and in order to compare the actual data from 2014 with the data from this study, we recreated maps from Collins and Amrein-Beardsley (2014) using the raw data available in that particular study that we reclassified into more general bins for comparative purposes (i.e., to more easily compare the data, then and now; see Figure 5).

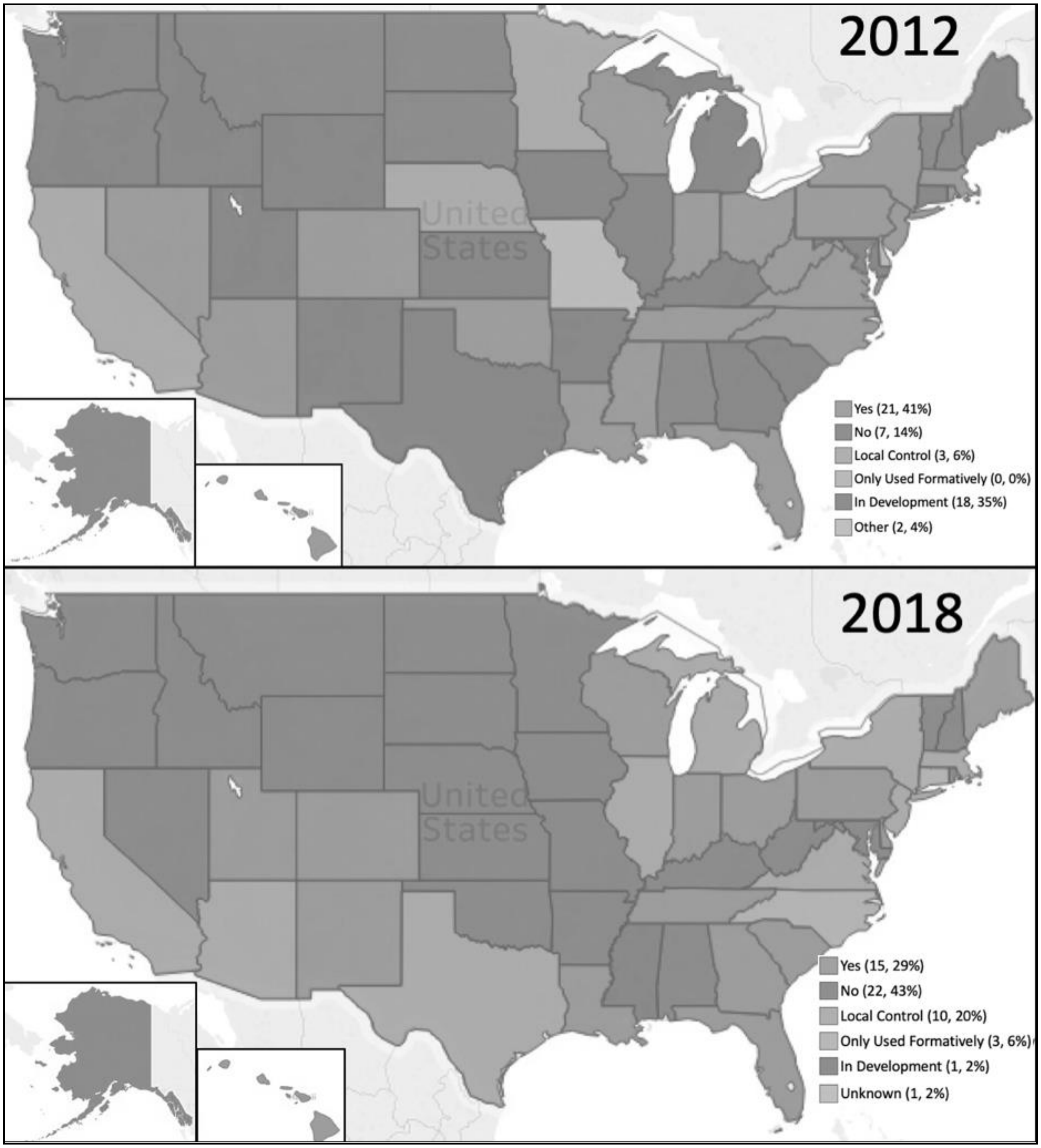
2012
Yes (21, 41%)
No (7, 14%)
Local Control (3, 6%)
Only Used Formatively (0, 0%)
In Development (18, 35%)
Other (2, 4%)
2018
Yes (15, 29%)
No (22, 43%)
Local Control (10, 20%)
Only Used Formatively (3, 6%)
In Development (1, 2%)
Unknown (1, 2%)
United States

Figure 5. Comparison map showing states that include VAMs as part of their teacher evaluation systems (2012, as per Collins & Amrein-Beardsley, 2014, to 2018).

Note: The number of states using VAMs decreased from 21 to 15 (41% to 29%, which is a decrease of 29%). The number of states not using VAMs increased from 7 to 22 (from 14% to 43% of states, which is an increase of 314%). The number of states reporting local control increased from 3 to 10 (from 6% to 20%, which is an increase of 333%). The number of states using VAMs only formatively increased from zero to three (from **0% to 6% of states,** which is an increase of 300%). The number of states with VAMs in development decreased from 18 to one (from 35% to 2% of states, which is a decrease of 94%). Lastly, the number of states classified as "other" decreased from two to one (from 4% to 2%, which is a decrease of 50%).

Most important to note from Figure 5 is that the number of states using state-wide VAMs decreased since 2012 (as per Collins & Amrein-Beardsley, 2014) from 41% to 29% of states (i.e., a decrease of 29%). Related, and perhaps more notably, the number of states that explicitly *do not* use or encourage VAM use substantially increased from seven of 51 states (14%) to 22 of 51 states (43%; i.e., an increase of 314%). Another important result of note is that in 2012 many states were still developing or piloting VAMs (including the aforementioned SGP), but in 2018 many of these states reversed their former VAM plans and trajectories. More specifically, in 2012, 18 of 51 states (35%) were piloting or developing a VAM; yet, in 2018 only the state of Mississippi (2%) reported having a VAM in development (i.e., a decrease of 94%). Additionally, the number of states that now leave decisions about VAM use to local school districts has increased from three to 10 (6% to 20%; i.e., an increase of 333%). This also demonstrates a substantial, and perhaps anticipated change, post–ESSA's (2016) shift toward more local control.

Additional state-by-state details regarding VAMs, as per this research question, can be found in Table 2. Again, this online table includes information about the types of assessments and grade level included in states VAMs (if used), the consequences attached to VAMs (if used), and the percentage of teachers' evaluation score for which VAMs are to count (if used) for both in 2012 (as per Collins & Amrein-Beardsley, 2014) and 2018. Likewise, Table 2 is an extension of Table 1, but it also includes state-by-state information from 2012 and 2018 side-by-side so that readers can compare specifics that are too detailed and space exhaustive to include along with these general results.

Research Question 3: Perceived Strengths and Weaknesses of States' Post–ESSA Systems

The following section includes an explanation of the perceived strengths and weaknesses of states' post–ESSA (2016) teacher evaluation systems (i.e., for states for which state personnel responded to this part of the study). Recall that 39 (76%) personnel from states' departments of education responded to the survey in total. Of those 39 individuals, 36 (71%) responded to this part of the survey and only 22 (44%) were willing to discuss their states' weaknesses. We aggregated these data to come up with broad themes protect the anonymity of their state sources, hence no illustrative maps revealing state by state data.

Strengths and weaknesses. The two overarching themes regarding strengths were increased stakeholder input in the process and increased formative feedback in the process. In terms of weaknesses, four overarching themes were evident. State department personnel were concerned that there was too much variety among teacher evaluation systems. Related, personnel were concerned that there was not enough capacity to support such variety and that there was a dearth of communication between states and local educational authorities (e.g., districts). Additionally, some personnel felt that the language of official policies should change to reflect a different attitude towards teachers (See Table 3).

Table 3
List of Overarching Themes and Prevalence Regarding Strengths and Weaknesses

	Theme	**Prevalence**
Strengths	Local control	24/36 (67%)
	Formative feedback	12/36 (33%)
Weaknesses	Too much variety	7/22 (32%)
	Not enough capacity to support	5/22 (23%)
	Lack of communication between states and districts	5/22 (23%)
	Need new language to reflect philosophical changes	5/22 (23%)

Strengths. For strengths, one major theme reflected the increased local control supposedly provided by ESSA. A majority of state department respondents (24 of 36; 67%) presented increased stakeholder inputs as their new systems' primary strengths. This was a common theme as per the results in both sections above regarding increased local control. Department personnel identified increased stakeholder inputs, particularly at the local level, as the primary factor that also helped to change and improve relationships between teachers and other education leaders and authorities (e.g., from "combative to cooperative").

Less prevalent, but still widely evident in the data (12 of 36; 33%) many state department personnel indicated that systems meant to be more collaborative than punitive were a strength of their states' post–ESSA (2016) teacher evaluation systems. These respondents emphasized the collaborative nature of their post–ESSA systems noting, more specifically, that they built their new systems with their conceptualizations of and understandings about how their states' teachers are to be evaluated in a new and, perhaps, reformed light. Instead of employing tools for measurement as imposed in authoritative manners, respondents noted that a strength of their new teacher evaluation systems are, again, meant to be collaborative and also help teachers improve their pedagogical practices via professional development and training.

Weaknesses. As for weaknesses, or areas for improvement across states' teacher evaluation systems, 22 of 36 (61%) state department personnel provided feedback. Maybe paradoxically, seven of these personnel (7 of 22; 32%) revealed difficulties with the sheer variety of teacher evaluation systems created by local school districts now causing states difficulties when conducting comparisons of districts within and across their states. More specific concerns in this area (5 of 22; 23%) included the extent to which personnel (on behalf of their respective states) felt that they might be able to provide policy and system support on a state-level scale (e.g., interpreting data from multiple albeit unfamiliar and unique district systems). Also, state department personnel (5 of 22; 23%) considered communication and contact points with local school districts to be an area of weakness regarding teacher evaluation systems. This could include, for example, improvements of states websites and the teacher evaluation information made public online and states' communication systems for training and support regarding states' teacher evaluation systems.

Lastly, other personnel (5 of 22; 23%) in this group reported that their states' teacher evaluation system language does not often match their new philosophies, policies, and general takes on their states' approaches to teacher evaluation. For example, statements explicating that states' systems are now meant to be more formative than summative are missing, as are broad statements about how ranking teachers as "ineffective" does not contribute to the philosophies and intentions underlying states' new teacher evaluation systems. In other words, these state department personnel would like to change the language or the content in official policies to include or reflect more about the intention of evaluation systems to help teachers learn, not to punish teachers.

Research Question 4: How Have Perceived Strengths and Weaknesses of States' Systems Changed Post–ESSA

Collins and Amrein-Beardsley (2014) included a similar set of questions posed to state department personnel about the perceived strengths and weaknesses of their states' teacher evaluation system prior to ESSA; hence, below are some key results also pertinent to state differences between now and then.

In 2012 (i.e., post–Race to the Top, 2011) the main concerns expressed by state personnel regarding their states' teacher evaluation systems largely pertained to issues with assessing student progress in non-tested areas (i.e., fairness, as described prior), general validity (as defined prior), and challenges with or desires to use the models formatively (versus summatively, which was the primary intent written into Race to the Top, 2011 and the NCLB waivers the federal government put into place around the same time, as also explained prior). Inversely, state department personnel in 2012 cited system strengths such as having comparable scores across districts (given states were federally incentivized to have uniform teacher evaluation systems at the time), having similar scores for core teachers across their states, having more measures for evaluating teachers (which were largely noted as the teacher-level observation systems described prior; see also Kane & Staiger, 2012), and having more "predictive power" (see also predictive validity described prior[7]) regarding future student success, again, as largely based on VAMs.

In 2018, state personnel's strength and weakness responses centered around the seemingly changed perceptions and intentions of states, as made explicit via states' post– ESSA (2016) teacher evaluation systems. Namely, that states are now to allow for more formative feedback to help teachers improve upon their pedagogical practices, more collaboration, and more stakeholder input and feedback (e.g., in the development, execution, and refinement of states' systems). However, while some state department personnel lauded increased communication between teachers and training offered to teachers, other state department personnel warned and worried that more local control meant less capacity for state departments to support diverse and multifarious teacher evaluation systems (e.g., in terms of providing districts support, training, appropriate communication systems, and appropriate quality controls). This was clearly evidenced as a policy and practice conundrum. A related issue, for example, was the extent to which states are now permitting districts to use multiple assessments to measure student growth, in varied ways, but also the extent to which districts understand how important it is to have the assessments that they adopt and use validated for their intended purposes. This is also, now more than prior, a noteworthy challenge (see also Sloat, Amrein-Beardsley, & Holloway, 2018).

The notable shifts in responses pre– and post–ESSA indicate states have taken more holistic views of and approaches towards their teacher evaluation systems, especially in comparison to the relatively more objective teacher evaluation systems in place prior. States' teacher evaluation policies and systems encourage more flexibility in practice, given multiple ways of measuring teacher effectiveness (also given the competing strengths and weaknesses of those additional measures). Put more simply, among state department personnel, there has been a profound change in how state leaders and personnel are talking and thinking about teacher evaluation post–ESSA.

Conclusions

We addressed what states' teacher evaluation systems look like post–ESSA (2016) and how states' teacher evaluation systems were in 2012 post–Race to the Top (2011) as compared to now (i.e., how states' teacher evaluation systems have changed over this 2012-2016 period of time of significant education policy enactment). While the purpose of this study was not to discover the underlying causes of such a complex shift in teacher evaluation systems in the US, researchers can infer the role that predominantly federal policies have played and continue to play in the state-level policies reviewed herein and prior. Rather, the purpose of this study was to provide an overview of data related to all states' teacher evaluation systems before and after the passage of ESSA (2015), especially because the rhetoric of ESSA may not match the actual policies.

---

[7] Predictive power, defined herein as essentially equivalent with predictive validity is evidenced when VAM-based and other teacher effectiveness estimates are used to predict future outcomes on a related academic (Kane, 2013; see also Kane, McCaffrey, Miller, & Staiger, 2013) or non-academic measure (e.g., lifetime earnings, pregnancy; see also Chetty, Friedman, & Rockoff, 2014a, 2014b)

First, VAMs are still in use as a component of teacher evaluation systems, but they are losing traction among state departments of education. This general trend is clear as per the data presented herein, as well as what would likely be expected after ESSA (2015) loosened the reigns on the federal incentives tied to states' use(s) of states' formerly reformed teacher evaluation models. More specifically, while some states continue to use VAMs, they do not include them as parts of the teacher evaluation scores or processes nearly as often, for nearly as much weight if still used, and definitely not nearly as often for high-stakes, consequential purposes. Instead, if VAMs are still being encouraged or used, they are being used to yield data which teachers might use to understand and then improve upon their own pedagogy and practice, as best they can (e.g., given some of the transparency and formative use issues with using VAMs, as discussed prior, are still at play). The implication of this finding is that VAMs may still play an important role in new wave of teacher evaluation systems, despite some belief that the passage of ESSA may eliminate VAMs. However, post-ESSA teacher evaluation systems which continue to use VAMs, overall, have reduced the weight of such measures in teachers' overall evaluation and have reduced or removed consequences tied to VAMs.

Second, the Danielson and Marzano observational frameworks seem to now be driving much of the action across teacher evaluation systems across the US, as likely related to the renewed and formative values and intentions clearly inherent in states' post–ESSA teacher evaluation systems. Such observational systems align better with states' new and apparent enthusiasms for teacher evaluation systems bent on formative use is also clear as per the evidence collected herein. This is also in line with recent research about effective teacher evaluation practices (Reinhorn, Moore Johnson, & Simon, 2017). Hence, we are starting to see a shift away from quantitative test score measures towards measures using scores from research-based conceptual frameworks like the Danielson or Marzano frameworks which break the complex activity of teaching into scored subcomponents meant to be used for formative purposes (e.g., discussion and professional development). The implication here is that policymakers or practitioners working on teacher evaluation systems in the current era should consider these additional evaluation frameworks, or at minimum, recognize the additional subcomponents that can be factored into teacher evaluation data.

Third, while there is still a legacy of emphases on VAMs as student growth measures, the definition of student growth is changing as well. In 2012, student growth essentially referred to growth as measured by states' standardized assessments of student achievement, aggregated, and then attributed to students' teachers' effects (e.g., as measured via VAMs). In 2018, student growth now includes other, more diverse, multiple measures, still including observational systems but also now including student surveys and SLOs. Put differently, the underlying construct (i.e., student growth) is the same, but the ways of defining and measuring it are different, more custom-made, and more holistic, given ESSA.

Fourth, there is a heightened emphasis on local control post–ESSA (2015) across states. While state department personnel expressed concerns about efficiently training and supporting local school districts with a large variety of systems, states have apparently responded to ESSA (2015), by-and-large, by allowing districts within their states to create what are essentially endorsed, curated, or completely homegrown teacher evaluation systems that can be customized to local school districts' desires, philosophies, and needs. State practices in this area unquestionably walk the line between manageability and flexibility. However, such practices may also set precedent for future teacher evaluation systems by providing both flexibility and support to local districts in the future. Additional research should consider whether local control creates a better environment for navigating the practical challenges of creating and implementing a teacher evaluation system. We recommend that policymakers continue to monitor how the heightened emphasis on local control plays out with regards to teacher evaluation systems.

Finally, the myriad lawsuits filed between teacher unions and state departments of education over the last decade (*Education Week*, 2015; see also Amrein-Beardsley & Close, 2019) may have driven some of the philosophical changes noted, especially in terms of more cooperative and formative, and less punitive and consequential teacher evaluation systems. Some state department personnel cited that new teacher evaluation systems with a focus on stakeholder involvement even changed teachers' and state leaders' relationships from "combative to cooperative." Perhaps this new era of teacher evaluation even reflects an honest effort to correct some of the pugnaciousness of the previous federal policies.

What is ultimately evidenced: ESSA has impacted the ways in which states are thinking about and enacting or endorsing teacher evaluation systems that do look different now than they did post–Race to the Top (2011). The reversal of trends, many would argue, constitute steps in the right direction, though those who still believe in high-stakes accountability systems at the teacher level, or student- or school-level may argue these steps are in the wrong direction. Regardless of stance, any persons interested in or concerned about the current state of states' teacher evaluation systems post–ESSA (2016) should have data, via this study, to understand changes these systems over time, at minimum. This should, accordingly, be of historical but also timely "value-added."

References

Alvarez, M. E., & Anderson-Ketchmark, C. (2011). Danielson's framework for teaching. *Children & Schools*, *33*(1), 61-63. https://doi.org/10.1093/cs/33.1.61

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, 44*(8), 448-452. https://doi.org/10.3102/0013189X15618385

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

American Statistical Association (ASA). (2014)._ASA statement on using value-added models for educational assessment. ASA. Retrieved from http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher, 37*(2), 65-75. https://doi.org/10.3102/0013189X08316420

Amrein, A. L. & Berliner, D. C. (2002). High-Stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1-74. Retrieved from https://doi.org/10.14507/epaa.v10n18.2002

Amrein-Beardsley, A., & Close, K. (2019). Teacher-level value-added models (VAMs) on trial: Empirical and pragmatic issues of concern across five court cases. *Educational Policy,* 1-42. https://doi.org/10.1177/0895904819843593

Bailey, K. M. (2001). Observation. In Carter, R., & Nunan, D. (Eds.), *Teaching English to speakers of other languages* (pp. 114-119). Cambridge University Press. https://doi.org/10.1017/CBO9780511667206.017

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

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., … Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute. Retrieved from http://www.epi.org/publications/entry/bp278

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. https://doi.org/10.14507/epaa.v21n5.2013

Barnett, J. H., Rinthapol, N., & Hudgens, T. (2014). *TAP research summary: Examining the evidence and impact of TAP. The System for Teacher and Student Advancement*. National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED556331.pdf

Betebenner, D. W. (2009). *Growth, standards and accountability.* The National Center for the Improvement of Educational Assessment, Inc. Retrieved from: http://www.nciea.org/publication_PDFs/growthandStandard_DB09.pdf

Betebenner, D. W. (2011). *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.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on Value-Added models.* Testing Service. Retrieved from www.ets.org/Media/Research/pdf/PICVAM.pdf

Braun, H. (2015). The value in value-added depends on the ecology. *Educational Researcher, 44*(2), 127-131. https://doi.org/10.3102/0013189X15576341

Capitol Hill Briefing. (2011). *Getting teacher evaluation right: A challenge for policy makers.* Dirksen Senate Office Building (Research in brief).

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Measuring the impact of teachers I: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2593-2632. https://doi.org/10.3386/w19424

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Measuring the impact of teachers II: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632. https://doi.org/10.3386/w19424

Chiang, H., McCullough, M., Lipscomb, S., & Gill, B. (2016). Can student test scores provide useful measures of school principals' performance? U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/2016002/pdf/2016002.pdf

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, 22(98).* https://doi.org/10.14507/epaa.v22.1594

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 16*(1). Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17291

Danielson, C. (2012). Observing classroom practice. *Educational Leadership*, *70*(3), 32-37.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* ASCD.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of 'No Child Left Behind'. *Race Ethnicity and Education*, *10*(3), 245-260. https://doi.org/10.1080/13613320701503207

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

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267-297. https://doi.org/10.1002/pam.21818

Eckert, J. M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Phi Delta Kappan, 91*(8), 88-92. https://doi.org/10.1177/003172171009100821

Education Week. (2015). *Teacher evaluation heads to the courts*. Retrieved from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

Every Student Succeeds Act (ESSA) of 2015, Pub. L. No. 114-95, § 129 Stat. 1802. (2015). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability and disability: Gaming the system. *Advances in Applied Microeconomics*, *14*, 35-49. https://doi.org/10.1016/S0278-0984(06)14002-X

Gabriel, R., & Lester, J. N. (2013). Sentinels guarding the grail: Value-added measurement and the quest for education reform. *Education Policy Analysis Archives, 21*(9), 1-30. https://doi.org/10.14507/epaa.v21n9.2013

Geiger, T. J., & Amrein-Beardsley, A. (2017). Administrators gaming test- and observation-based teacher evaluation methods: To conform to or confront the system. *American Association of School Administrators (AASA) Journal of Scholarship and Practice, 14*(3), 45-53.

Geiger, T. J., & Amrein-Beardsley, A. (in press). Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys. *Cogent Education.* https://doi.org/10.1080/2331186X.2019.1602943

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value-added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*(2), 96-104. https://doi.org/10.3102/0013189X15575031

Graue, M. E., Delaney, K. K., & Karch, A. S. (2013). Ecologies of education quality. *Education Policy Analysis Archives, 21*(8), 1-36. https://doi.org/10.14507/epaa.v21n8.2013

Green, P. C., Baker, B. D., & Oluwole, J. (2012). Legal and policy implications of value-added teacher assessment policies. *The Brigham Young University Education and Law Journal*, *1.*

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/10.1177/0395909808320678

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives, 8*(41). https://doi.org/10.14507/epaa.v8n41.2000

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

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

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

Ho, A. D., Lewis, D. M., & MacGregor Farris, J. L. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(4), 15-26. https://doi.org/10.1111/j.1745-3992.2009.00159.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* MET Project. Bill & Melinda Gates Foundation. Retrieved from https://files.eric.ed.gov/fulltext/ED540959.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

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. https://doi.org/10.14507/epaa.v23.1968

Kelley, K., Clark, B., Brown, V., & Sitzia, J. (2003). Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, *15*(3), 261-266. https://doi.org/10.1093/intqhc/mzg031

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

Koretz, D. (2017). *The testing charade: Pretending to make schools better.* The University of Chicago Press. https://doi.org/10.7208/chicago/9780226408859.001.0001

Kraft, M. A, & Gilmour, A. F. (2017). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5) 234-249. https://doi.org/10.3102/0013189X17718797

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). *How states use student learning objectives in teacher evaluation systems: A review of state websites* (REL 2014-013). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory North-east & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=380

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis, 38*(4), 738–756. https://doi.org/10.3102/0162373716666166

Marzano, R. J. (2012). The two purposes of teacher evaluation. *Educational Leadership*, 70(3), 14-19.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement.* ASCD.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101. https://doi.org/10.3102/10769986029001067

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. SAGE.

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

National Institute for Excellence in Teaching. (NIET). (n.d.). Retrieved from https://www.niet.org/

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23), 1-27. https://doi.org/10.14507/epaa.v18n23.2010

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

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2002). Retrieved from http://www.ed.gov/legislation/ESEA02/

Perry, J., & Vogell, H. (2009, October 19). Are drastic swings in CRCT scores valid? *Atlanta Journal-Constitution* . Retrieved from www.ajc.com/news/news/local/are-drastic-swings-in-crct-scores-valid/nQYQm

Qualtrics [Computer software]. (2019). Retrieved from http://www.qualtrics.com

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

Reinhorn, S. K., Moore Johnson, S., & Simon, N. S. (2017). *Educational Evaluation and Policy Analysis*, *39*(3), 383–406. https://doi.org/10.3102/0162373717690605

Rhee, M. (2011, April 6). The evidence is clear: Test scores must accurately reflect students' learning. *Huffington Post* . Retrieved, from www.huffingtonpost.com/michelle-rhee/michelle-rhee-dc-schools_b_845286.html

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247-252. https://doi.org/10.1257/0002828041302244

Ross, E., & Walsh, K. (2019). *State of the States 2019: Teacher and Principal Evaluation Policy*. National Council on Teacher Quality.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, (*4*)*4*, 537-571. https://doi.org/10.3386/w14666

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 175-214. https://doi.org/10.1162/qjec.2010.125.1.175

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET Project*. National Education Policy Center (NEPC). Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299-311. https://doi.org/10.1007/bf00973726

Sanders, W. L. (2003, April). *Beyond "No Child Left Behind."* Paper presented at the annual meeting of the American Educational Research Association, Chicago. Retrieved February 10, 2007, from http://www .sas.com/govedu/edu/no-child.pdf

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS EVAAS*. SAS Institute Inc.

SAS Institute Inc. (n.d.). *SAS EVAAS for K-12: Assess and predict student performance with precision and reliability*. Author.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142-171. https://doi.org/10.3102/1076998611432174

Shaw, L. H. & Bovaird, J. A. (2011, April). *The impact of latent variable outcomes on value-added models of intervention efficacy*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

Sloat, E., Amrein-Beardsley, A., Holloway, J. (2018). Different teacher-level effectiveness estimates, different results: Inter-model concordance across six generalized value-added models (VAMs). *Educational Assessment, Evaluation and Accountability, 30*(4), 367-397. https://doi.org/10.1007/s11092-018-9283-7

Sloat, E., Amrein-Beardsley, A., & Sabo, K. E. (2017). Examining the factor structure underlying the TAP System for Teacher and Student Advancement. *AERA Open, 3*(4), 1–18. https://doi.org/10.1177/2332858417735526

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability:

Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340-359. https://doi.org/10.1162/EDFP_a_00186

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

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.

Tableau [Computer software]. (2019). Retrieved from http://www.tableau.com

Toppo, G., Amos, D., Gillum, J., & Upton, J. (2011). When test scores seem too good to believe. *USA Today*. Retrieved from www.usatoday.com/news/education/2011-03-06-school-testing_N.htm

Upton, J. (2011). For teachers, many ways and reasons to cheat on tests. USA Today . Retrieved from www.usatoday.com/news/education/ 2011-03-10-1Aschooltesting10_CV_N.htm

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

U.S. Department of Education (USDOE). (2012). *Elementary and Secondary Education Act (ESEA) flexibility.* Author. Retrieved from https://www.ed.gov/esea/flexibility

U.S. Department of Education (USDOE). (2014). *States granted waivers from No Child Left Behind allowed to reapply for renewal for 2014 and 2015 school years.* Author. Retrieved from http://www.ed.gov/news/press-releases/states-granted-waivers-no-child-left-behind-allowed-reapply-renewal-2014-and-2015-school-years

U.S. Department of Education (USDOE). (n.d.). *Targeting growth using student learning objectives as a measure of educator effectiveness.* Author. Retrieved Nov. 11, 2018 from https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/targeting-growth.pd f

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in place: How new teacher evaluations fail to live up to promises.* National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsView/Final_Evaluation_Paper

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* The New Teacher Project (TNTP). Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

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

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. https://doi.org/10.1016/j.econedurev.2009.07.004

Yeh, S. S. (2013). A re-analysis of the effects of teacher replacement using value-added modeling. *Teachers College Record, 115*(12), 1-35. Retrieved from http://www.tcrecord.org/Content.asp?ContentID=16934

About the Authors

Kevin Close
Arizona State University
Email: kevin.close@asu.edu
ORCID: https://orcid.org/0000-0003-1643-5124
Kevin Close is currently pursuing a PhD in the Learning, Literacies, and Technologies program at Arizona State University. His research focused on digital adaptive assessments, nation-wide teacher evaluation systems based on high-stakes tests, and design in education. His interests lie in using technology to change the way we assess and measure progress.

Audrey Amrein-Beardsley
Arizona State University
Email: audrey.beardsley@asu.edu
ORCID: https://orcid.org/0000-0002-1250-2281
Audrey Amrein-Beardsley, PhD., is a Professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the use of value-added models (VAMs) in and across states before and since the passage of the Every Student Succeeds Act (ESSA). More specifically, she is conducting validation studies on multiple system components, as well as serving as an expert witness in many legal cases surrounding the (mis)use of VAM-based output.

Clarin Collins
Arizona State University
Email: clarin.collins@asu.edu
ORCID: https://orcid.org/0000-0003-1630-9881
Clarin Collins, Ph.D., is Director of Scholarly Initiatives in the Mary Lou Fulton Teachers College at Arizona State University. Her research interests include national and state policy implementation at the local level, teacher interaction with and influence on education policy, and education accountability and evaluation systems.

About the Guest Editor

Audrey Amrein-Beardsley
Arizona State University
audrey.beardsley@asu.edu
Audrey Amrein-Beardsley, PhD., is a Professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the use of value-added models (VAMs) in and across states before and since the passage of the Every Student Succeeds Act (ESSA). More specifically, she is conducting validation studies on multiple system components, as well as serving as an expert witness in many legal cases surrounding the (mis)use of VAM-based output.

SPECIAL ISSUE
Policies and Practices of Promise in Teacher Evaluation

education policy analysis archives

Volume 28 Number X Month day, 2020 ISSN 1068-2341

Readers are free to copy, display, distribute, and adapt this article, as long as the work is attributed to the author(s) and **Education Policy Analysis Archives**, the changes are identified, and the same license applies to the derivative work. More details of this Creative Commons license are available at https://creativecommons.org/licenses/by-sa/2.0/. **EPAA** is published by the Mary Lou Fulton Institute and Graduate School of Education at Arizona State University Articles are indexed in CIRC (Clasificación Integrada de Revistas Científicas, Spain), DIALNET (Spain), Directory of Open Access Journals, EBSCO Education Research Complete, ERIC, Education Full Text (H.W. Wilson), QUALIS A1 (Brazil), SCImago Journal Rank, SCOPUS, SOCOLAR (China).

Please send errata notes to Audrey Amrein-Beardsley at audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.

education policy analysis archives
editorial board

Lead Editor: **Audrey Amrein-Beardsley** (Arizona State University)
Editor Consultor: **Gustavo E. Fischman** (Arizona State University)
Associate Editors: **Melanie Bertrand, David Carlson, Lauren Harris, Eugene Judson, Mirka Koro-Ljungberg, Daniel Liou, Scott Marley, Molly Ott, Iveta Silova** (Arizona State University)

Cristina Alfaro
San Diego State University
Gary Anderson
New York University
Michael W. Apple
University of Wisconsin, Madison

Jeff Bale
University of Toronto, Canada
Aaron Bevenot SUNY Albany

David C. Berliner
Arizona State University
Henry Braun Boston College

Casey Cobb
University of Connecticut
Arnold Danzig
San Jose State University
Linda Darling-Hammond
Stanford University
Elizabeth H. DeBray
University of Georgia
David E. DeMatthews
University of Texas at Austin
Chad d'Entremont Rennie Center for Education Research & Policy
John Diamond
University of Wisconsin, Madison
Matthew Di Carlo
Albert Shanker Institute
Sherman Dorn
Arizona State University
Michael J. Dumas
University of California, Berkeley
Kathy Escamilla
University ofColorado, Boulder
Yariv Feniger Ben-Gurion University of the Negev
Melissa Lynn Freeman
Adams State College
Rachael Gabriel
University of Connecticut

Amy Garrett Dikkers University of North Carolina, Wilmington
Gene V Glass
Arizona State University
Ronald Glass University of California, Santa Cruz

Jacob P. K. Gross
University of Louisville
Eric M. Haas WestEd

Julian Vasquez Heilig California State University, Sacramento
Kimberly Kappler Hewitt
University of North Carolina Greensboro
Aimee Howley Ohio University

Steve Klees University of Maryland
Jaekyung Lee SUNY Buffalo
Jessica Nina Lester
Indiana University
Amanda E. Lewis University of Illinois, Chicago
Chad R. Lochmiller Indiana University
Christopher Lubienski Indiana University
Sarah Lubienski Indiana University

William J. Mathis
University of Colorado, Boulder
Michele S. Moses
University of Colorado, Boulder
Julianne Moss
Deakin University, Australia
Sharon Nichols
University of Texas, San Antonio
Eric Parsons
University of Missouri-Columbia
Amanda U. Potterton
University of Kentucky
Susan L. Robertson
Bristol University

Gloria M. Rodriguez
University of California, Davis
R. Anthony Rolle
University of Houston
A. G. Rud
Washington State University

Patricia Sánchez University of University of Texas, San Antonio
Janelle Scott University of California, Berkeley
Jack Schneider University of Massachusetts Lowell
Noah Sobe Loyola University

Nelly P. Stromquist
University of Maryland
Benjamin Superfine
University of Illinois, Chicago
Adai Tefera
Virginia Commonwealth University
A. Chris Torres
Michigan State University
Tina Trujillo
University of California, Berkeley
Federico R. Waitoller
University of Illinois, Chicago
Larisa Warhol
University of Connecticut
John Weathers University of Colorado, Colorado Springs
Kevin Welner
University of Colorado, Boulder
Terrence G. Wiley
Center for Applied Linguistics
John Willinsky
Stanford University
Jennifer R. Wolgemuth
University of South Florida
Kyo Yamashiro
Claremont Graduate University
Miri Yemini
Tel Aviv University, Israel

archivos analíticos de políticas educativas
consejo editorial

Editor Consultor: **Gustavo E. Fischman** (Arizona State University)
Editores Asociados: **Felicitas Acosta** (Universidad Nacional de General Sarmiento), **Armando Alcántara Santuario** (Universidad Nacional Autónoma de México), **Ignacio Barrenechea, Jason Beech** (Universidad de San Andrés)**,** **Angelica Buendia**, (Metropolitan Autonomous University), **Alejandra Falabella** (Universidad Alberto Hurtado, Chile), **Carmuca Gómez-Bueno** (Universidad de Granada), **Veronica Gottau** (Universidad Torcuato Di Tella), **Carolina Guzmán-Valenzuela** (Universidade de Chile), **Antonia Lozano-Díaz** (University of Almería), **Antonio Luzon**, (Universidad de Granada), **María Teresa Martín Palomo** (University of Almería), **María Fernández Mellizo-Soto** (Universidad Complutense de Madrid), **Tiburcio Moreno** (Autonomous Metropolitan University-Cuajimalpa Unit), **José Luis Ramírez,** (Universidad de Sonora), **Axel Rivas** (Universidad de San Andrés), **César Lorenzo Rodríguez Uribe** (Universidad Marista de Guadalajara), **Maria Veronica Santelices** (Pontificia Universidad Católica de Chile)

Claudio Almonacid
Universidad Metropolitana de Ciencias de la Educación, Chile

Miguel Ángel Arias Ortega
Universidad Autónoma de la Ciudad de México

Xavier Besalú Costa
Universitat de Girona, España

Xavier Bonal Sarro Universidad Autónoma de Barcelona, España

Antonio Bolívar Boitia
Universidad de Granada, España

José Joaquín Brunner Universidad Diego Portales, Chile

Damián Canales Sánchez
Instituto Nacional para la Evaluación de la Educación, México

Gabriela de la Cruz Flores
Universidad Nacional Autónoma de México

Marco Antonio Delgado Fuentes
Universidad Iberoamericana, México

Inés Dussel, DIE-CINVESTAV, México

Pedro Flores Crespo Universidad Iberoamericana, México

Ana María García de Fanelli
Centro de Estudios de Estado y Sociedad (CEDES) CONICET, Argentina

Juan Carlos González Faraco
Universidad de Huelva, España

María Clemente Linuesa
Universidad de Salamanca, España

Jaume Martínez Bonafé
Universitat de València, España

Alejandro Márquez Jiménez
Instituto de Investigaciones sobre la Universidad y la Educación, UNAM, México

María Guadalupe Olivier Tellez, Universidad Pedagógica Nacional, México

Miguel Pereyra Universidad de Granada, España

Mónica Pini Universidad Nacional de San Martín, Argentina

Omar Orlando Pulido Chaves
Instituto para la Investigación Educativa y el Desarrollo Pedagógico (IDEP)

José Ignacio Rivas Flores
Universidad de Málaga, España

Miriam Rodríguez Vargas
Universidad Autónoma de Tamaulipas, México

José Gregorio Rodríguez
Universidad Nacional de Colombia, Colombia

Mario Rueda Beltrán Instituto de Investigaciones sobre la Universidad y la Educación, UNAM, México

José Luis San Fabián Maroto
Universidad de Oviedo, España

Jurjo Torres Santomé, Universidad de la Coruña, España

Yengny Marisol Silva Laya
Universidad Iberoamericana, México

Ernesto Treviño Ronzón
Universidad Veracruzana, México

Ernesto Treviño Villarreal
Universidad Diego Portales Santiago, Chile

Antoni Verger Planells
Universidad Autónoma de Barcelona, España

Catalina Wainerman
Universidad de San Andrés, Argentina

Juan Carlos Yáñez Velazco
Universidad de Colima, México

arquivos analíticos de políticas educativas
conselho editorial

Editor Consultor: **Gustavo E. Fischman** (Arizona State University)
Editoras Associadas: **Andréa Barbosa Gouveia** (Universidade Federal do Paraná), **Kaizo Iwakami Beltrao**, (Brazilian School of Public and Private Management - EBAPE/FGVl), **Sheizi Calheira de Freitas** (Federal University of Bahia), **Maria Margarida Machado**, (Federal University of Goiás / Universidade Federal de Goiás), **Gilberto José Miranda**, (Universidade Federal de Uberlândia, Brazil), **Marcia Pletsch** (Universidade Federal Rural do Rio de Janeiro), **Maria Lúcia Rodrigues Muller** (Universidade Federal de Mato Grosso e Science), **Sandra Regina Sales** (Universidade Federal Rural do Rio de Janeiro)

Almerindo Afonso
Universidade do Minho
Portugal

Rosanna Maria Barros Sá
Universidade do Algarve
Portugal

Maria Helena Bonilla
Universidade Federal da Bahia
Brasil

Rosa Maria Bueno Fischer
Universidade Federal do Rio Grande do Sul, Brasil

Alice Casimiro Lopes
Universidade do Estado do Rio de Janeiro, Brasil

Suzana Feldens Schwertner
Centro Universitário Univates
Brasil

Geovana Mendonça Lunardi Mendes Universidade do Estado de Santa Catarina

Flávia Miller Naethe Motta
Universidade Federal Rural do Rio de Janeiro, Brasil

Alexandre Fernandez Vaz
Universidade Federal de Santa Catarina, Brasil

Regina Célia Linhares Hostins
Universidade do Vale do Itajaí, Brasil

Alfredo Macedo Gomes
Universidade Federal de Pernambuco
Brasil

Jefferson Mainardes
Universidade Estadual de Ponta Grossa, Brasil

Jader Janer Moreira Lopes
Universidade Federal Fluminense e Universidade Federal de Juiz de Fora, Brasil

Debora Nunes
Universidade Federal do Rio Grande do Norte, Brasil

Alda Junqueira Marin
Pontifícia Universidade Católica de São Paulo, Brasil

Dalila Andrade Oliveira
Universidade Federal de Minas Gerais, Brasil

José Augusto Pacheco
Universidade do Minho, Portugal

Jane Paiva
Universidade do Estado do Rio de Janeiro, Brasil

Paulo Alberto Santos Vieira
Universidade do Estado de Mato Grosso, Brasil

Fabiany de Cássia Tavares Silva
Universidade Federal do Mato Grosso do Sul, Brasil

António Teodoro
Universidade Lusófona
Portugal

Lílian do Valle
Universidade do Estado do Rio de Janeiro, Brasil

Alfredo Veiga-Neto
Universidade Federal do Rio Grande do Sul, Brasil