GETTING DOWN TO FACTS II

Technical Report

In Need of Improvement? Assessing the California Dashboard after One Year

Morgan S. Polikoff
Shira Korn
Russell McFall

University of Southern California

September 2018

About: The *Getting Down to Facts* project seeks to create a common evidence base for understanding the current state of California school systems and lay the foundation for substantive conversations about what education policies should be sustained and what might be improved to ensure increased opportunity and success for all students in California in the decades ahead. *Getting Down to Facts II* follows approximately a decade after the first Getting Down to Facts effort in 2007. This technical report is one of 36 in the set of *Getting Down to Facts II* studies that cover four main areas related to state education policy: student success, governance, personnel, and funding.

Stanford University

PACE
Policy Analysis for California Education

In Need of Improvement? Assessing the California Dashboard after One Year

Morgan S. Polikoff
Shira Korn
Russell McFall

University of Southern California

California is taking a bold new approach to the evaluation of school performance, the provision of data to parents and educators, and the use of accountability to improve student learning. Broadly speaking, the state is stepping away from the consequential accountability that was compelled under No Child Left Behind and toward a system that relies more on capacity building. The state is also moving away from a single measure to evaluate school performance (tossing the Academic Performance Index (API) on the dustbin of history) and moving toward a much more complex, multi-faceted approach. Owing to a more hands-off US Department of Education under more recent accountability legislation, the state is charting its own course.

The so-called "California Way" is the subject of great discussion in Sacramento and the nation's capital, and there are some promising signs that it is off to a good start (Johnson & Tanner, 2018). In general, the "California Way" consists of:

- Dramatic increases in educational spending and redistribution of state funds toward school districts serving more disadvantaged districts (see the reports on LCFF and the LCAPs for more information on this);
- A sharp move away from consequential ("rewards and sanctions") accountability and toward a "continuous improvement" approach to addressing low performance;
- A strong stand in favor of multiple measures of school performance rather than just one (typically A-F or 0-100) rating, as used in more than 40 other states.

The centerpiece of the second and third bullets is the new California School Dashboard, a website that gives school and district leaders and parents evidence about the performance of California schools. We describe the Dashboard in detail below.

The purpose of this report is to provide early evidence about how well the system is working. To do so, we draw on a four-part framework for appraising accountability systems first put forth by Polikoff, McEachin, Wrabel, and Duque (2014). This framework identifies construct validity, reliability, fairness, and transparency as key features by which to judge accountability systems. We explain these criteria and introduce a new feature—consequences—and apply them to the California School Dashboard to evaluate the progress of the new system.

We ask three research questions about the early implementation of the California School Dashboard and whether stakeholders (i.e., superintendents and parents) are using it as intended.

1. To what extent does the dashboard align with what we know about effective accountability systems?
2. In its early implementation, how are superintendents using the dashboard? What do they say about its strengths and weaknesses? How do their uses align (or not) with the theory of action for the policy?
3. What do parents think of the dashboard and multiple measures accountability? How do opinions differ according to key demographic groups?

We draw on three data sources to answer these questions. To answer the first question, we analyze statewide dashboard data drawn from the CDE website. To answer the second question, we draw on data from statewide interviews of superintendents conducted as a part of the broader Getting Down to Facts project. To answer the third question, we use data from the 2018 PACE/USC Rossier state-representative poll of California voters.

The remainder of the report proceeds as follows. First, we briefly summarize the literature on school accountability, including both its intended and unintended consequences. Second, we present our framework for analyzing school accountability systems. Third, we describe our data and analytic approach. Fourth, we answer our three broad research questions. We conclude by discussing next steps for research around the dashboard.

Background

The Effects of Accountability[1]

Because accountability has been a prominent feature of state and federal education policy over the last several decades, there has been a substantial literature on its implementation and effects (see Figlio & Loeb, 2011, for a more complete treatment). This large literature has three main areas of emphasis: effects on students, effects on teachers, and unintended consequences. We note at the outset that virtually all of the results we describe in this section come from studies of consequential accountability policies—the extent to which these results generalize to softer, Dashboard-style accountability systems is not clear.

Student outcomes. The literature on the effects of accountability policies on student outcomes generally finds that these policies can boost student achievement on both high- and low-stakes assessments. This evidence comes from cross-state studies that leverage variation in the timing of adoption of standards policies (i.e., Dee & Jacob, 2011; Wong, Cook, & Steiner, 2015) as well as within-state longitudinal studies (e.g., Jacob, 2005; Klein, Hamilton, McCaffrey, & Stecher, 2000) and studies that compare schools on either side of an accountability threshold (e.g., Ahn & Vigdor, 2014; Bonilla & Dee, 2017; Chakrabarti, 2014). In general, these effects appear to be larger for low-performing schools (Figlio & Rouse, 2006; Jacob, 2005), in states with higher proficiency standards (Wong et al., 2015), and in states with greater degrees of local autonomy (Loeb & Strunk, 2007). These effects might be considered modest in magnitude, but the fact that they accrue to all students suggests the total impact is impressive.

Three other relevant student outcomes that have been investigated are achievement gaps, non-test long-term outcomes, and student engagement. Studies of the effects of accountability on achievement gaps suggest that, while these policies have led to some degree of gap closure, these gap closure effects appear quite modest (e.g., Dee & Jacob, 2011; Figlio, Rouse, & Schlosser, 2009; Gaddis & Lauen, 2014; Lauen & Gaddis, 2012). The one study we are aware of that investigates long-term outcomes found positive effects of accountability pressure on college enrollment, completion, and earnings (Deming, Cohodes, Jennings, & Jencks, 2016).

[1] Portions of this section are adapted from Polikoff and Korn (in press).

In contrast the possibility of rewards for high achievement led to meaningful reductions in long-term outcomes, as schools were able to reclassify low-performing students as disabled, showing that the design of accountability policies clearly matters. Finally, a recent study of the impact of NCLB on survey measures of student engagement found that the policy caused an initial positive bump, but that over time the impact of NCLB on engagement became negative (Markowitz, 2018).

Teacher outcomes. A key desired outcome of accountability policies is instructional change, and there is reasonably strong evidence that these policies can indeed drive teachers to modify their instruction. Research makes clear that teachers *believe* they devote significant effort to aligning instruction with state standards (e.g., Hamilton & Berends, 2006; Pedulla et al., 2003). Observational research finds, however, that instruction remains moderately aligned to standards and often reflects a poor understanding of the expectations embedded in standards (e.g., Cohen, 1990; Hill, 2001; McDonnell, 2004; Spillane, 2006). Sophisticated survey studies show that: 1) alignment of instruction to standards is modest (Polikoff, 2012a); 2) alignment to standards has increased gradually over time, particularly in mathematics (Polikoff, 2012a); and 3) some policies and teacher characteristics (e.g., better-aligned standards and assessments, more experienced teachers) predict better instructional alignment (Polikoff, 2012b, 2013). There is also a growing literature on the effects on teacher working conditions that finds these effects are mixed and quite modest, contradicting popular claims that accountability has made teaching less attractive (e.g., Dee et al., 2013; Grissom, Nicholson-Crotty, & Harrington, 2014; Reback et al., 2011; Sun, Saultz, & Yee, 2017).

Other unintended consequences. Finally, there is quite a large literature on the unintended consequences of accountability for both students and teachers. Teachers and schools under accountability sometimes strategically target their efforts so as to artificially maximize their scores and avoid accountability consequences. For instance, there is evidence of "teaching to the test" by reallocating instructional time to tested subjects (e.g., Dee et al., 2013; Hamilton et al., 2005; Judson, 2013) or teaching specific test-taking strategies (e.g., taking practice tests, learning how to complete specific test items, or focusing on content that is expected to be assessed) (Jennings & Bearak, 2014; Jennings & Sohn, 2014). Teachers may also attempt to game accountability systems by targeting resources and instructional efforts on students who matter the most for a schools' rating (e.g., the so-called "bubble students") (e.g., Booher-Jennings, 2005; Jennings & Sohn, 2014). School leaders sometimes engage in harmful gaming behaviors, such as reallocating the "least effective" teachers to non-tested grades (Grissom et al., 2017). Finally, there have been instances of outright cheating related to accountability tests; in addition to news reports of explicit cheating scandals, one study in Chicago estimated that approximately 5% of teachers had cheated on the state exams (Jacob & Levitt, 2003). It is important to note that, though these behaviors can be pernicious, evidence of the impacts of accountability on no-stakes NAEP exams suggests that these gaming behaviors probably do not explain all of the achievement effects.

Designing Accountability Systems

To analyze the dashboard, we rely on a four-part framework first used by Polikoff and colleagues (2014) to examine states' NCLB waiver policies. This framework rates states' policies on their construct validity, reliability, transparency, and fairness. We add a fifth element to the framework—consequences—for reasons we describe below.

Construct validity. In the context of accountability system design, construct validity refers to two issues. First, to what extent do the performance measures cover the latent set of desired student outcomes? In other words, do the measures the state uses to evaluate schools adequately capture all of the outcomes policymakers believe are important? This leads to the second issue, are the inferences we draw about school effectiveness from these measures appropriate? In other words, if we conclude that a school is high or low performing, are we right? Naturally, if we do not include all relevant measures of performance in our accountability metric, our school ratings will not accurately depict school performance.

There is general acceptance that schools should be judged for more than just student test scores. As such, the NCLB and waiver accountability systems generally fared poorly on construct validity as these systems typically relied exclusively or heavily on math and ELA proficiency as the key metric of school performance, choosing a very narrow slice of what it means to be a "good school." Furthermore, the heavy reliance on proficiency rates and other status-based measures of performance meant that these systems typically did a poor job identifying schools based on how effective they were at improving student learning.[2]

For the purpose of this paper, we examine construct validity largely theoretically, discussing the measures in the dashboard and whether they fit the definition of a construct valid system as just discussed. We also consider the extent to which the measures relate to one another (converge or diverge) as an indicator of the extent to which they measure their underlying constructs (for example, we expect math and reading performance levels to be highly positively correlated and we expect math performance levels and math performance growth to be modestly positively correlated given how those constructs are related in the population).

Reliability. In multiple measures accountability systems, reliability refers to the extent to which the measures are consistent over time[3]. For instance, reliable measures of school

[2] The state is not very explicit about the intended goals of the Dashboard system; we describe what we can glean from materials on the state website in the sections below. If the state is primarily concerned about identifying schools that have students who need additional support, then a status-based system may well be appropriate. We infer from state materials that the state is, to a large extent, interested in identifying schools that are particularly effective or ineffective (we also note that the state plans to use the Dashboard to identify schools for consequential accountability under ESSA). If this is true, then metrics based on status will be less appropriate for the reasons described here and below. In short, the intended goals of the system matter quite a lot for the judgment of construct validity.

[3] If the state were reporting multiple measures of the same thing – multiple achievement status measures, for instance – and aggregating them in a single index, then we might also ask to what extent the measures are correlated with one another. Since the state explicitly does not do this, we only look at year-to-year consistency.

performance will tend to correlate highly with themselves from year to year, sending consistent messages about schools' performance. If a system is not reliable (the same measure fluctuates over time), this can send a confusing message to schools about performance. NCLB accountability metrics based on percent proficient were highly reliable. In general, growth-based measures of performance will be less reliable, but one way to improve their reliability is to use multiple-year averages, as some states did under the NCLB waivers. In this paper, we examine reliability primarily by considering the year-to-year correlation of dashboard measures.

Transparency. Transparency refers to the extent to which the plan is documented and understandable by its consumers (largely parents and educators). If stakeholder groups cannot make sense of the accountability system, almost by definition it will be unable to drive performance improvements. State accountability systems under NCLB were fairly transparent because they relied on one, relatively easy to understand measure. However, there were many hidden complications that actually made the system much less transparent than it seemed on the surface (see e.g., Davidson, Reback, Rockoff, & Schwartz, 2015). While including more measures may present a more holistic assessment of school progress, it might also be hard for parents and administrators to interpret, especially in schools that perform well in some areas but not in others. In this analysis we discuss transparency through argument, but we also present evidence on the number of measures available and the types and frequency of missing data.

Fairness. The fairness question for accountability systems measures to what extent schools are assessed on factors beyond their control. In other words, when schools are being evaluated for their effectiveness, are those evaluations biased by things like student demographics or school size or location, factors that are presumably out of the schools' control? Fairness is closely related to construct validity, but measures could certainly be high on one dimension but not another. For instance, a measure could be fair in that it is uncorrelated with student demographics, but it could also be completely irrelevant for educational effectiveness (something like height might fit here). Or a set of measures could be highly construct valid—broad and covering a wide range of desirable outcomes—but not very fair—biased against more diverse schools. Note that this definition of fairness is based on fairness to *schools*, not fairness to *students*; the latter might point in different directions.

NCLB accountability systems were largely considered unfair, as their reliance on proficiency rates as the sole measures of school effectiveness meant that schools serving low achieving populations often performed poorly on their ratings regardless of their contributions to student learning. This kind of status-based measure of performance is highly correlated with student demographics and very weakly correlated with schools' actual contributions to student learning. Here, we evaluate fairness by considering the correlation of dashboard measures with school-level demographic variables (e.g., percent from various racial/ethnic groups, percent receiving free or reduced-price lunch) to assess whether student demographic characteristics are strong predictors of school ratings. We acknowledge that this is a flawed test, as it is possible (indeed, probable) that school quality/effectiveness is in fact unevenly distributed. We discuss the limitations of our approach below, and we do not foreground fairness concerns in our implications section.

Consequences. The four-part framework was used to analyze new systems before they were implemented. However, the dashboard is a year into implementation, so a fifth and final important aspect to evaluate is its consequences. We examine evidence of the consequences of the new system and especially consider the extent to which the consequences of the accountability system align with its intended consequences (Reckase, 1997). The goals of NCLB accountability were to shine light on poor performance and, concomitantly, raise achievement and narrow achievement gaps. NCLB's consequences, as described above, did drive improvements in instruction and student achievement, but they also led to undesired unintended consequences. Though only one year under the dashboard has elapsed, we can begin to get a sense for the consequences by examining the responses of district leaders and parents to the new dashboard system—do their responses to interview and survey questions suggest that the system is on track to have its desired consequences? We describe what we believe to be these desired consequences next.

What is the California School Dashboard?

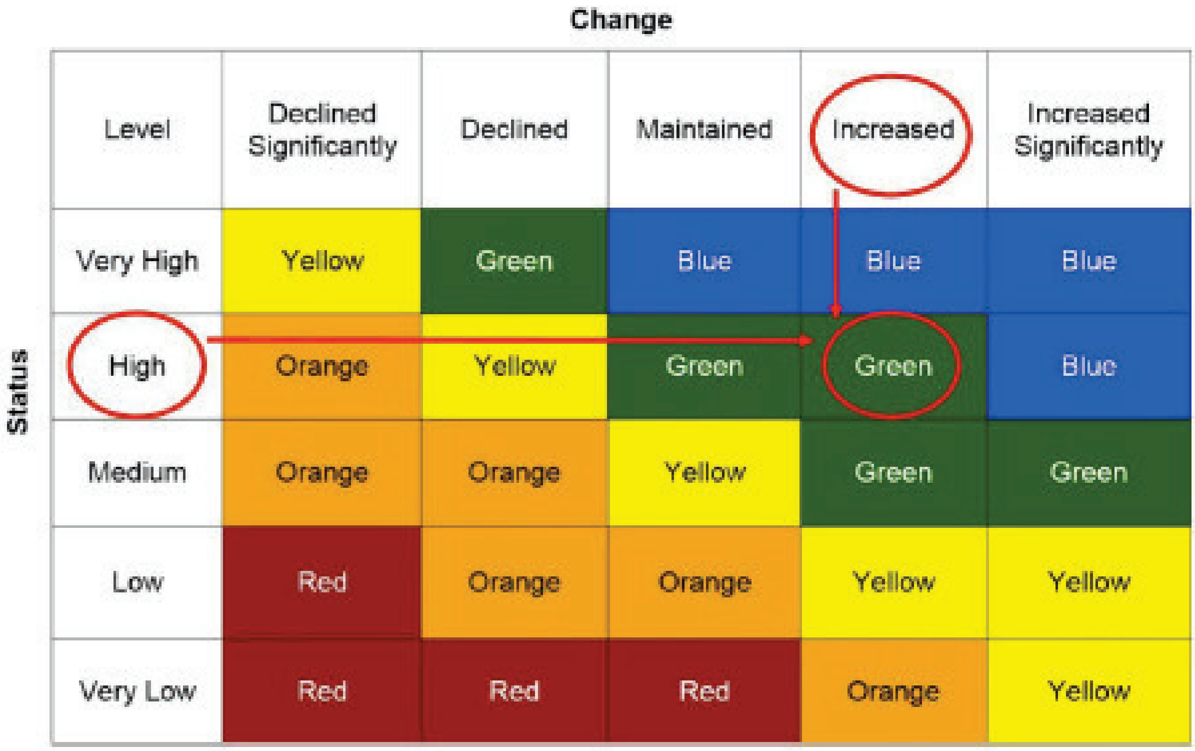
The origins of the Dashboard. Under NCLB, California had parallel accountability systems. The NCLB measure, Adequate Yearly Progress (AYP), was used for NCLB's consequential accountability. The state measure (enacted under the Public Schools Accountability Act), was the Academic Performance Index (API). This 200-to-1000 measure was widely used around the state by parents and even realtors as the shorthand for school quality. The presence of two accountability systems was widely seen as confusing and unnecessary. When given the opportunity to consolidate systems, California took advantage.

The California School Dashboard is California's consolidated accountability system. The Dashboard fulfills the accountability requirements of the Local Control Funding Formula (LCFF) and the Every Student Succeeds Act (ESSA). The LCFF has eight district priorities, which range from pupil achievement to parent engagement. Districts are required to describe how they will address each of these priorities in their Local Control and Accountability (LCAP) plans. The measures in the dashboard provide data for five of these eight district priorities. Thus, the Dashboard is intended to 1) Support Local Education Agencies (LEAs) in identifying strengths, weaknesses, and areas for improvement; 2) Assist in determining whether LEAs are eligible for assistance; and 3) Assist the State Superintendent for Public Instruction in determining whether LEAs are eligible for more intensive state support/intervention[4].

Under ESSA, the state must develop a method for identifying low-performing schools for intervention. Here, the Dashboard is intended to serve as the method for identifying schools for ESSA accountability. For more information on the intention and history of the dashboard see (California Department of Education [CDE], 2017).

[4] Another purpose for the Dashboard might be to inform the public about school performance. This does not seem to be a major intended focus of the Dashboard as we discuss somewhat below.

The form of the Dashboard. The Dashboard, available at www.caschooldashboard.org, contains detailed ratings for California public schools (for details on which schools are included and excluded, see CDE, 2017a). Each school's Dashboard report presents school performance on an array of indicators. For non-high schools, these are chronic absenteeism, suspension rate, English learner (EL) progress, English language arts achievement, and mathematics achievement. For high schools, the achievement measures are replaced with a graduation rate measure and a college/career readiness measure. For each indicator, the Dashboard provides information on four main areas: equity report, status and change report, detailed report, and student group report.



Change

	Level	Declined Significantly	Declined	Maintained	Increased	Increased Significantly
Status	Very High	Yellow	Green	Blue	Blue	Blue
	High	Orange	Yellow	Green	Green	Blue
	Medium	Orange	Orange	Yellow	Green	Green
	Low	Red	Orange	Orange	Yellow	Yellow
	Very Low	Red	Red	Red	Orange	Yellow

Figure 1. Example performance level calculation. This school earned a score of "high" for its status level and "increased" for its change level, giving it an overall green rating.

The equity report provides the overall school ratings on each of the indicators. It also provides a count of the number of subgroups for which there is reporting on each indicator, as well as the number of subgroups receiving a low score (orange or red) on that indicator. The status and change report again is focused on overall school ratings, but it provides numerical and categorical scores for each indicator. This is the "behind the scenes" on how the overall scores are calculated. The detailed report provides some longitudinal data about the various indicators. Finally, the student group report shows how each student group rates on each indicator. Clicking on any indicator pulls up a page that provides even more detail about performance, overall and for each subgroup, on that indicator.

Performance for each indicator is determined through a combination of current performance (Status) and improvement over time (Change). For each indicator, the school's current performance level is first assigned a rating of very high, high, medium, low, or very low, as seen on the vertical axis of Figure 1. The school's change level (which is calculated by merely subtracting last year's score on the indicator from this year's score) is assigned a rating of increased significantly, increased, maintained, decreased, or decreased significantly, as seen on the horizontal axis of Figure 1. Based on the status and change classifications, and according to a figure like the one shown above in Figure 1, schools are then given a color-coded performance level (red, orange, yellow, green, or blue). On the Dashboard website, schools and districts are given these color-coded performance levels for each of the indicators, both overall and for each numerically significant student group.

The goals of the Dashboard. The state is not especially explicit about the goals of the Dashboard in its ESSA plan or in the Dashboard documentation, but there is some evidence in press releases and on the state Department of Education website as to the theory of action behind the Dashboard. For example, the "core messages" the state puts forth regarding the Dashboard are: 1) that students are more than a test score and multiple measures are necessary to diagnose strengths and weaknesses of a school; 2) that the Dashboard highlights inequities to help communities improve the standing of low-performing student groups; 3) that improvement must be a local endeavor and the Dashboard should provide the information for local decisionmakers to use; and 4) that support, assistance, and continuous improvement (not punishment) are how California schools will improve (CDE, 2017b). Based on these messages, we can infer that the state thinks school effectiveness is multi-faceted; that the primary role for improving low-performing schools lies with local actors; and that rewards and sanctions are not a core part of the "California way."

There are other common goals of accountability systems that California's Dashboard does not seem to share. For instance, some systems use rewards and sanctions to incentivize better performance. California's leaders clearly do not believe that approach is right for the state. Another goal of an accountability system might be to provide parents with information to compare schools and inform their choices. Again, this does not seem to be a major focus of the Dashboard, at least as currently described in state materials. It is important to keep these intentions in mind when we present our results, especially those pertaining to consequences.

Data

There are three data sources for this analysis: school-level CA Dashboard data[5], downloaded from the California Department of Education website (which we augment with school-level demographic and other variables from the Common Core of Data); interviews of 91

[5] Ideally we would have access to student-level data which would allow us to conduct additional analyses such as evaluating the relationship between the state's "change" measures and true value-added measures. We are unable to conduct such analyses, however, given that we only have access to school-level data. There are serious concerns about the change measures and what they measure, which we discuss throughout this report.

California district superintendents; and statewide poll data from the 2018 PACE/USC Rossier poll of California registered voters.

School-level Dashboard Data

Our main data source for the analyses presented below is the Fall 2017 CA Dashboard data that are used to populate the Dashboard website. We downloaded, cleaned, and merged the Dashboard datasets for each of the Dashboard indicators. We exclude school districts from all analyses, reporting only on school-level Dashboard data. We describe any Dashboard variables we use as we present the results below.

Superintendent Interviews

We add to the evidence from the Dashboard by drawing on evidence from a qualitative dataset containing superintendents' responses to a series of interview questions about the Dashboard. These questions are shown in Figure 2.

1. What do you see as one or more of the main purpose(s) of the Dashboard (e.g., how money should be spent, what equity means)
2. How have you used the new Dashboard?
    a. What metrics have you used from the Dashboard?
3. Are there metrics that you think are missing?
    a. If so, which ones?
    b. Why?
4. What metrics do you use (if any) to learn about schools in your district that you are concerned about?
5. Do you feel pressure to improve any one of the metrics in particular?
    a. If yes, where does this pressure come from?

Figure 2. Questions about the California Dashboard asked in superintendent interviews

These questions were included in a broader interview protocol that also covered topics such as personnel, district finances, and teaching and learning in the district. The final superintendent interview sample contains 91 interviews from a stratified random sample of 205 (a response rate of 44%). We generally rely on the pre-existing codes that were done by interview team members (for more on the sampling and coding, see the work by Moffitt and colleagues in this series). In some instances, we also coded the available interview responses on additional dimensions—we describe these coding efforts below as we present results from the interviews.

PACE/USC Rossier Poll

Finally, we included a series of questions about the Dashboard on the 2018 PACE/USC Rossier poll of registered California voters (for more details on the poll, see Polikoff, 2018). This online, state-representative poll was administered in late January of 2018 to 2500 registered

California voters. It focused on a range of California education policy issues, including charter schools, the Local Control Funding Formula, and the Dashboard/multiple measures accountability. The specific questions we asked related to the Dashboard were:

- Q16 (split sample with Q17)[6]. Some states give each school a single overall performance rating, usually on an A-F or 0-100 scale. Other states do not give schools a single rating, and instead give each school an individual rating on several different measures of performance such as student assessment in specific subject areas, graduation rates, and suspension rates. Which of these best expresses your opinion about an overall rating?
    - I think schools should receive an overall rating, such as on an A-F or 0-100 scale.
    - I do not think schools should receive an overall rating, but instead they should receive a rating for multiple measures of performance.
    - Don't know.
- Q17 (split sample with Q16) Under the new federal education law, more than 40 states have chosen to give each of their schools a single overall performance rating, usually on an A-F or 0-100 scale. California has chosen not to give schools an overall rating; instead, schools will be rated on several different measures but the scores will not be combined. Which of these best expresses your opinion about an overall rating?
    - I think schools should receive an overall rating, such as on an A-F or 0-100 scale.
    - I do not think schools should receive an overall rating, but instead they should receive a rating for multiple measures of performance.
    - Don't know.
- Q18. In 2017, California changed the way it evaluated schools. The new California Dashboard evaluates school performance on multiple measures. To what extent are you familiar with the new California Dashboard?
    - I have heard or read about it, and I know a good deal about it.
    - I have heard or read about it, and I know a fair amount.
    - I have heard or read about it, but I do now know much about it.
    - I have never heard or read about it.
    - Don't know.
- Q19 (asked only of those who reported having heard of the Dashboard in Q18). Have you visited the new California Dashboard website?
    - No, I have never visited.
    - Yes, I have visited once or twice.
    - Yes, I have visited several times/more.
    - Don't know.
- Q20 (asked only of those who reported having heard of the Dashboard in Q18). Based on what you know, do you have a positive or negative impression of the California Dashboard?
    - Very positive

[6] We randomly split the sample for two reasons. First, we did not know the ideal phrasing of this question that would elicit the most reliable information. Second, we were interested in this survey experiment to test the impact of question framing on respondents' views.

- Somewhat positive
- Somewhat negative
- Very negative
- Don't know

- Q21 (after showing respondents two images of the Dashboard). Based on what you see here, do you have a positive or negative impression of the California Dashboard?
  - Very positive
  - Somewhat positive
  - Somewhat negative
  - Very negative
  - Don't know
- Q22. Please indicate how much you agree or disagree with the following statements about the California School Dashboard. For each one, please indicate whether you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with that policy. If you are not sure, please indicate so.
  - The new California School Dashboard captures the most important measures of district and school quality and performance.
  - The information displayed on the new California School Dashboard is easy to understand.
  - The information displayed on the new California School Dashboard is an effective means to communicate outcomes to the community.

In the results section, we analyze these questions, presenting both topline and crosstab results.

Results

Construct Validity

We examine the construct validity of the Dashboard system with both data from the Dashboard and the superintendent interviews. Specifically, we assess construct validity through the following questions: 1) To what extent do the set of performance measures adequately cover the latent set of student outcomes state leaders desire?; And 2) Are the inferences the state makes about school effectiveness on the basis of those measures appropriate?

Covering the latent set of desired outcomes. There is no question that, when compared with the API system it replaces, the Dashboard does a much better job covering the latent set of desired outcomes of California schools. Whereas AYP and the API were focused more-or-less exclusively on standardized test score levels in mathematics and ELA, the Dashboard evaluates schools on a variety of measures. In addition to test score levels, the Dashboard covers outcomes such as:

- Test score growth (more on the quality of the "change" indicators as measures of actual growth below).
- A measure of attainment – the four-year adjusted cohort graduation rate and its change relative to the previous three years.

- The suspension rate.
- A composite measure of "college and career readiness" that includes CTE pathway completion, state assessments (Smarter Balanced), AP and IB exams, success in dual enrollment, and completion of the A-G course requirements.
- The progress of English language learners (EL progress), a large and historically underserved group in California schools.[7]
- Chronic absenteeism (in future years).

This list is by no means exhaustive. We might also wish to include items such as more direct measures of culture/climate, measures of student opportunity and access (e.g., to quality teachers), achievement in subjects other than mathematics and ELA, and perhaps stronger measures of college and (particularly) career readiness that actually track long-term outcomes in our evaluation of schools. Nonetheless, we consider the array of measures included in the Dashboard a marked improvement from those included under the previous system.

In district interviews, superintendents were asked whether any metrics were missing from the Dashboard. Of the 91 superintendents, 38 said or implied that there were metrics that were missing (versus 31 who said or implied no and 22 who either were not asked the question[8] or provided some other answer). Among these respondents, 10 reported that evidence of college/career readiness was missing, 9 reported school climate, 7 reported social/emotional health, and 20 reported other measures were missing. Of those who indicated other measures were missing, examples included parent involvement, drilled-down information on college/career readiness, and local assessments. Several respondents also noted that they would like to be able to compare schools and districts with similar contexts (e.g., locations, sizes, demographics), and the current tool would not allow them to do that.

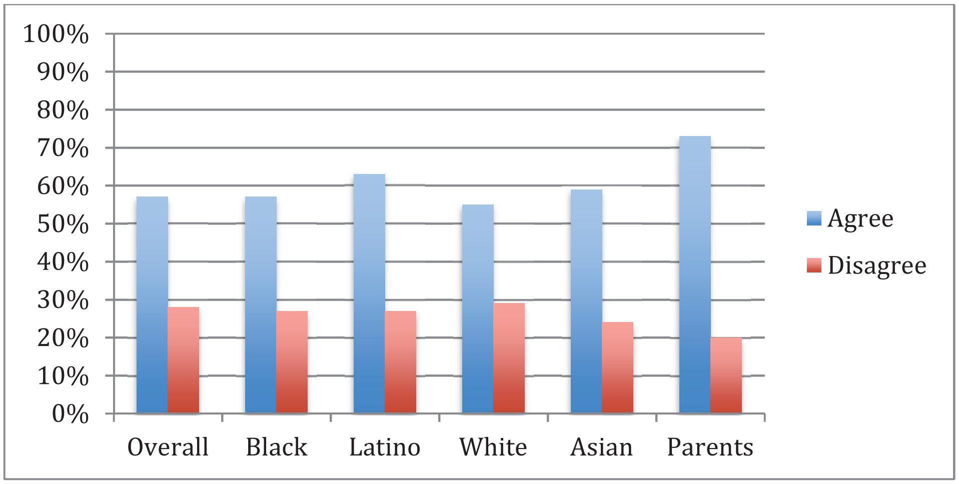


Figure 3. Voters' agreement with the statement "The new California School Dashboard captures the most important measures of district and school quality and performance."

[7] Progress is defined as the percent of eligible EL students who either a) advance one level from year to year on a six-level proficiency index on the CELDT test or b) test in the early advanced or advanced levels in both years.
[8] Since we did not conduct the interviews, we do not know why some respondents were not asked the question.

The PACE/USC Rossier poll also provides evidence about the completeness of the measures in the Dashboard. After seeing images of the Dashboard, voters were asked whether they agreed that it captured the most important measures; 57% agreed (12% strongly), versus just 28% who disagreed (7% strongly) as shown in Figure 3. This result held for all racial/ethnic groups, all regions of the state, and all education levels. Agreement was especially strong among parents, 73% of whom agreed (24% strongly) versus just 20% who disagreed (5% strongly). In short, California voters in general, and parents specifically, believe the Dashboard captures the most important measures. This is perhaps not surprising given that large majorities of voters report that it is important to judge schools on an array of measures beyond test scores, including social/emotional skills, art/music, and graduation/dropout rates.

Making appropriate inferences. As for whether the inferences made about school performance under the Dashboard are appropriate, there are both conceptual and empirical reasons to be skeptical. First, the Dashboard is still heavily (at least 50%) reliant on status measures of performance. For all of the indicators, schools that score in the lowest category on status cannot earn higher than a yellow rating, even if they earn very high growth ratings relative to other schools in the state. Given that research is clear that school effectiveness—the extent to which a school actually improves outcomes for students—is measured by growth and not status (see for instance Linn, 2003; Neal & Schanzenbach, 2010), this degree of emphasis on status certainly reduces the construct validity of the system. That said, the state did wisely choose to base status on average student scores, rather than on rates of proficiency, which researchers agree is the best that can be done with a status measure of performance (Polikoff, 2016).[9]

Furthermore, while there are many ways to construct growth models recommended by the literature, California uses a poor method. For instance, while the state has the data to calculate models based on individual student growth that more closely approximate the causal impact of schools (i.e., value-added models or, failing that, student growth percentiles), it instead calculates a crude growth measure by simply subtracting last year's mean achievement from this year's (even though many of the students in this year's data are likely to be different from those in last year's). Researchers advocate for a true value-added model or student growth percentile model because, among other reasons, these models account for compositional changes in the student body across years (whereas subtracting means does not). Furthermore, subtraction-based methods of calculating "growth" are conceptually inappropriate—the highest-performing school in the state in value-added terms might show no change in average achievement levels from year to year, because the value-added is already factored into each year's average score. It is not clear that these change measures actually measure "effectiveness" at all—their construct validity in that sense is quite weak, and recent analyses provided to the State Board of Education confirm that the current change measures are inaccurate in their identification most of the time (Hough & Miller, 2017). Amending the

[9] If the goal were simply to identify the schools with the students most in need of support, then the status measures are appropriate and well targeted (though in that case, such a complex system is surely not needed—the state could simply rank order schools by percent in poverty). This would, however, be an odd goal given the expense and effort designed to make a public-facing system like the Dashboard.

Dashboard calculations to reflect these methods is relatively straightforward and could be implemented at quite modest cost (basically, the cost to pay someone to calculate the value-added estimates).

Evidence from the Dashboard data. Correlations of Dashboard measures for elementary and middle schools are shown in Table 1, and correlations for high schools are shown in Table 2[10]. For both analyses, the Dashboard data confirms that the measures are generally correlated as expected[11]. Looking first at elementary and middle schools, we expect that schools that have high test scores will also have high EL progress and low suspension rates (because we expect that high performance is generally associated with low discipline problems). Indeed, we see that math and reading status are very highly correlated with each other (*r* = .94), and that both are moderately correlated with EL progress (*r* = .30 and *r* = .44, respectively). We also see that suspension rates are lower in schools with higher math and ELA status (*r* = -.44 and *r* = -.33, respectively). The only status correlation that might not match expectations is the slight positive correlation of EL progress with suspension rates (*r* = .15). Correlations of status with change measures for elementary schools are also as expected—all positive but modest. The lowest is the correlation of ELA status with ELA change (*r* = .14). The highest is the correlation of EL progress status with EL progress change (*r* = .35). In short, for the elementary grades, the patterns of relationships among the measures are in line with what we would expect given the relationships of the underlying constructs.

The patterns are also as expected at the high school level, as shown in Table 2. The measures based largely on test scores—ELA and math scores and the CCR indicator—are all correlated *r* =.77 or greater. These are also moderately positively correlated with graduation rates (in the *r* = .50 to *r* = .60 range) and somewhat less so with EL progress (in the *r* = .20 to *r* = .30 range). Again, suspensions are negatively correlated with the test-based measures, but their correlations with EL progress and graduation rates are quite close to zero. Finally, the measures are generally correlated as expected between status and change, with positive correlations in all cases except for graduation rates (where status and change are correlated *r* =-.04).

Overall, these results suggest that the Dashboard measures are behaving approximately as expected given what we would expect of the relationships among the underlying constructs. Two possible exceptions to this pattern are suspension rates at both elementary/middle and high school levels and graduation rate change scores. These might merit further investigation.

Taking all of the construct validity evidence together, do we think the Dashboard captures what we care about in evaluating schools? The Dashboard data themselves are certainly an improvement over API/AYP, and they clearly send an important signal that children

[10] There is a very small number of schools – approximately 160, or <2% of schools – where the school type variable does not correspond to the Dashboard indicators provided in the state data. These schools are excluded from all analyses.

[11] We do not respond statistical significance for two main reasons. First, we have essentially the population of schools in our data, so there is no need to talk about inferences from a sample to a population. Second, given the very large samples, almost all correlations in all tables would be statistically significant at conventional levels.

(and schools) are more than just test scores. However, there may be room for additional measures as the system matures. The "change" measures are extremely poor as indicators of actual student growth in achievement, and the state's poor design choices continue to conflate status and growth as measures of performance. Most school district leaders are reasonably content, but those who weren't offered some specific recommendations for important additions. And parents are generally content, although we did not specifically ask them if there were other measures they would like to see. Overall, the construct validity evidence appears middling, with several possibilities for improvement (we discuss implications in the final section).

Reliability

We examine the reliability of the Dashboard system with data from the Dashboard only. Specifically, we assess reliability through the following question: What is the year-to-year stability of the Dashboard measures?

Our ability to examine the reliability of Dashboard measures is limited because the current Dashboard only contains one year of "change" measures. Thus, we can only calculate year-to-year correlations for the status measures. As is typical for status measures (McEachin & Polikoff, 2012), all of the Dashboard measures are highly correlated from year to year. The weakest correlation is for EL progress, which is correlated $r = .62$ with last year's measure of EL progress. The rest of the year-to-year correlations range from $r = .83$ for suspension rates to $r = .97$ for math and ELA performance at grades 3-8. This indicates that, in general, schools receive similar ratings from year to year and the status measures in the Dashboard are highly reliable. The reliability of the change measures cannot be calculated due to the limited years of data on these measures, nor can the reliability of the color-coded levels. This is an important area for future investigation, as it could shed light on whether schools are making similar progress across years. Given what is known about the reliability of growth scores generally, we would expect that the reliability of change scores would be much lower than the reliability of status scores due to measurement error, changes in classroom characteristics, etc. (see for instance Kane & Staiger, 2002; Sass, 2008).

Transparency

We examine the transparency of the Dashboard system with data from the Dashboard, the superintendent interviews, and the PACE/USC Rossier poll. Specifically, we assess transparency through the following questions: 1) How many students and subgroups are included in the Dashboard accountability system and how does this compare to alternatives?; 2) What are superintendents' views of the transparency of the Dashboard? And 3) To what extent do voters report understanding the Dashboard?

Evidence from Dashboard data. The Dashboard system is more transparent about subgroup performance than California's previous accountability system by virtue of its smaller subgroup size. As a reminder, under the previous system, subgroup performance was reported for subgroups with n = 100 or greater, or if 50 ≤ n ≤ 100 and the subgroup made up 15% or more of the school population. Under the Dashboard, school performance is reported for any subgroup with 30 or more members.

As shown in Table 3, the Dashboard produces a sharp increase in the average number of subgroups that are reported. Under the old system, schools were required to report an average of 3.65 subgroups. Under the Dashboard, that number is 5.44 (about a 50% increase). The table also breaks out the increase for each subgroup. All subgroups see an increase in the numbers of schools that are required to report on them. However, the largest increases are for African American (from 14% of schools to 35%), Asian (from 18% to 38%), White (from 53% to 71%), multiple races (from 3% to 29%), English learners (from 62% to 82%), and students with disabilities (from 36% to 84%). Clearly, there have been dramatic increases in the number of subgroups for which schools are responsible.

Correspondingly, the Dashboard sharply reduces the number of students who are excluded from accountability because they are in schools where their subgroup size is too small, as shown in Table 4. For example, under the previous system 39% of California's African American students were in schools where the African American subgroup was not large enough to be reported; under the Dashboard this number is just 13%. For Asian students, the numbers are 24% under the old system and 8% under the new system. For Filipino students, the numbers are 64% under the old system and 31% under the new system. For multiple race students, the numbers are 80% under the old system and 28% under the new system. Finally, for students with disabilities, the numbers are 38% under the old system and 3% under the new system.

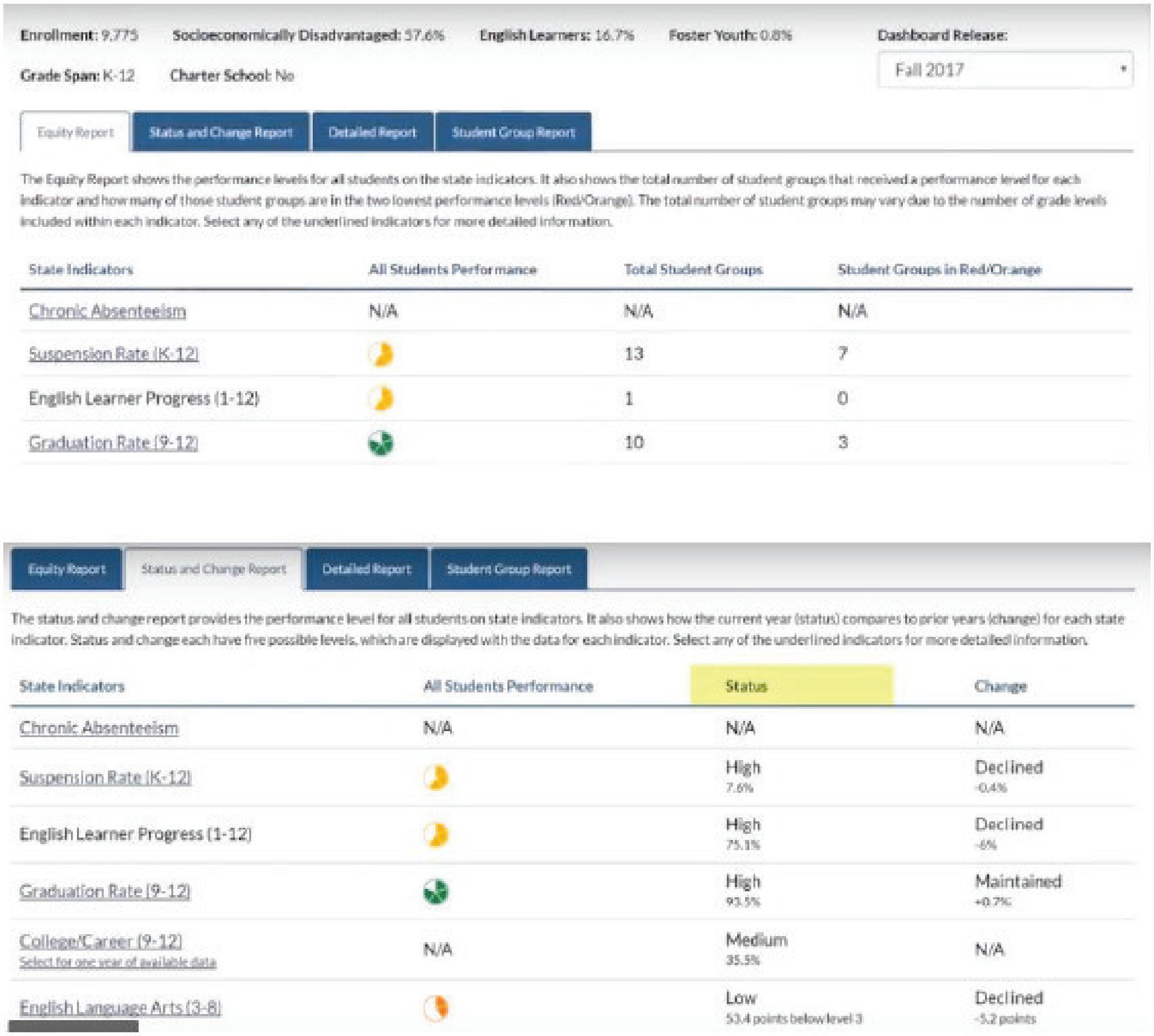
Of course, even more subgroups would be reported and students included if the minimum n-size were smaller. For instance, if the minimum n-size were reduced to 20 (as it is in Connecticut, among other states), the average number of reported subgroups would increase from 5.44 to 6.03, as shown in Table 3. The proportion of schools reporting African American students as a subgroup would increase to 43%, Asians would increase to 46%, Whites would increase to 78%, multiple races would increase to 40%, English learners would increase to 86%, and students with disabilities would increase to 90%. The percentage of students excluded from subgroup accountability would, correspondingly, decrease, as shown in Table 4. For African Americans, just 8% would be excluded; for Asians, just 5%; for Whites, just 1%; for multiple races, just 16%; for ELs, just 0.6%; and for students with disabilities, just 1%.

The main tradeoff to choosing a lower subgroup size in a typical accountability system would be that schools might be more subject to negative consequences just due to the random fluctuations in performance that are more likely in small subgroups. A second possible tradeoff is that smaller subgroups (and therefore more year-to-year fluctuations) might encourage educators to chase what is essentially measurement error in working to improve performance. That said, given California has virtually no consequential accountability and subgroup performance is mostly presented for transparency rather than consequences, it may make sense to explore lowering the subgroup size. Given that many other states use smaller n-sizes and choosing a smaller n-size would substantially increase the proportion of students included in subgroup accountability, California might consider lowering its n-size. We know of no literature that identifies a “right” n-size, but the state could explore alternatives and their impact on the Dashboard measures and then make a decision.

Evidence from superintendent interviews. In the interviews, superintendents provided evidence related to the transparency question. Instead of discussing which students were counted in the metrics, the superintendents raised concerns over missing or outdated data. First, 27 superintendents noted that the Dashboard was either incomplete (e.g., missing data) or not current, and that this affected the utility of the Dashboard for them. For example, several superintendents noted that data currently in the Dashboard dated from as far back as 2015, and that such old data could not be useful for them. Since we did not conduct the interviews (and therefore could not follow up), we don’t know what superintendents were referring to with these concerns; we did not see any old data in our look through the Dashboard, nor did we see missing data on a scale anywhere near 1-in-4 schools. Perhaps this perception of lack of currency contributed to the finding that 24 of the 91 principals (just over one-quarter) did not indicate they had used the data for any purpose. And of those who did report using the Dashboard data, many said their usage was limited because of either the incompleteness or the datedness of the data.

Evidence from the voter poll. Respondents to the poll were given the opportunity to see images of the California School Dashboard, and then all were asked whether “the information displayed on the new California School Dashboard is easy to understand” and “the information displayed on the new California School Dashboard is an effective means to communicate outcomes to the community.”

For the first question, 59% agreed (17% strongly) that the Dashboard is easy to understand, versus just 36% who disagreed (12%) strongly. Agreement on this question outweighed disagreement for every racial ethnic category, in every region of the state, and at every level of education. Parents were even more overwhelming in their agreement—74% agreed (28% strongly) that the Dashboard was easy to understand, versus just 24% who disagreed (7% strongly).



Enrollment: 9,775 **Socioeconomically Disadvantaged:** 57.6% **English Learners:** 16.7% **Foster Youth:** 0.8% **Dashboard Release:** Fall 2017

Grade Span: K-12 **Charter School:** No

The Equity Report shows the performance levels for all students on the state indicators. It also shows the total number of student groups that received a performance level for each indicator and how many of those student groups are in the two lowest performance levels (Red/Orange). The total number of student groups may vary due to the number of grade levels included within each indicator. Select any of the underlined indicators for more detailed information.

State Indicators	All Students Performance	Total Student Groups	Student Groups in Red/Orange
Chronic Absenteeism	N/A	N/A	N/A
Suspension Rate (K-12)	◕	13	7
English Learner Progress (1-12)	◕	1	0
Graduation Rate (9-12)	●	10	3

The status and change report provides the performance level for all students on state indicators. It also shows how the current year (status) compares to prior years (change) for each state indicator. Status and change each have five possible levels, which are displayed with the data for each indicator. Select any of the underlined indicators for more detailed information.

State Indicators	All Students Performance	Status	Change
Chronic Absenteeism	N/A	N/A	N/A
Suspension Rate (K-12)	◕	High 7.6%	Declined -0.4%
English Learner Progress (1-12)	◕	High 75.1%	Declined -6%
Graduation Rate (9-12)	●	High 93.5%	Maintained +0.7%
College/Career (9-12) Select for one year of available data	N/A	Medium 35.5%	N/A
English Language Arts (3-8)	◔	Low 53.4 points below level 3	Declined -5.2 points

Figure 4. Snapshot of the Dashboard that was shown to poll respondents

For the second question, 61% agreed (14% strongly) that the Dashboard is an effective means to communicate outcomes to the community, versus just 28% (8% strongly) who disagreed. Again this support held across racial ethnic groups, regions, and education levels. And support was even higher among parents, with 76% agreeing (26% strongly) versus 18% disagreeing (3% strongly). Despite complaints in the popular media and from pundits, parents appear to believe that the Dashboard is relatively transparent. Whether that belief carries over to understanding (i.e., do Californians who say the Dashboard is easy to understand actually understand it?) is an important question for future research.

Fairness

We operationalize fairness as the extent to which schools are accountable for things outside their control, such as the demographics of the students they enroll, their size, or their location. This has been an issue in previous systems; under NCLB, for example, larger schools or schools with more subgroups were much less likely to meet adequate yearly progress targets. This may be less of an issue under the Dashboard system because the negative consequences for poor performance are limited to nonexistent. To examine this question, we report correlations of status and change measures with various school characteristics obtained from the Common Core of Data, including demographic proportions (i.e., the percent of students falling under each demographic category), urbanicity, and school size (we use a log transformation as is typical). Rather than reporting all of the correlations, we summarize patterns, organized according to the outcome measure. We focus here on elementary and middle grades, where the data are more complete.

One limitation of this analysis is that we do not have an external measure of school effectiveness (e.g., a value-added estimate), so we cannot know whether correlations we observe are indicative of bias as opposed to a reflection of actual differences in the quality of education provided in schools that vary along these dimensions. That is, it may be case that California's urban schools are actually systematically better or worse than its non-urban schools. Thus, if we find a correlation between urbanicity and performance ratings, it may be that these are reflective of those effectiveness differences. Given this limitation, we provide the analysis below as suggestive but do not draw implications from this portion of the work.

Achievement status and change measures. Test-based status measures (e.g., math and ELA test scores) are very highly correlated with student demographic characteristics, as mentioned previously and shown elsewhere (see for instance McEachin & Polikoff, 2012). For example, the 1-to-5 level score associated with grades 3-8 math achievement is correlated -.60 with the percent of students who are Hispanic and -.78 with the percent of students who are socioeconomically disadvantaged. Both math and ELA test-based status measures are significantly correlated with every demographic proportion in the dataset. In contrast, they are only marginally correlated with location-based indicators (suburban schools have slightly higher math achievement, schools located in towns have slightly lower) and with total enrollment (larger schools have slightly higher math achievement (r = .04).

In contrast to the status measures, test-based change measures are weakly correlated with all of the demographic and school variables. For instance, the strong correlation of either math or ELA change scores with demographic variables is just r = .06. Schools serving more Asian students are slightly more likely to have higher Dashboard scores, and schools serving more students with disabilities are slightly less likely. Overall, the change scores do not seem to vary systematically between schools with different demographics. Thus, despite the imperfections of the change scores as a growth measure, this pattern suggests (but does not prove) that the fairness results presented above *do* reflect unfairness as opposed to actual differences in school effectiveness.

Suspension rate status and change measures. The patterns of association between suspension rate measures and school characteristics and student demographics are similar in direction but much weaker in magnitude. For instance, the suspension rate status measure is positively correlated with percent Asian (this measure is reverse coded, so higher values on the suspension rate measure are better) and negatively correlated with percent African American and percent socioeconomically disadvantaged, but these correlations are no higher than .27 (versus .80 for academic achievement measures). It would not be surprising for schools serving more low income, African American, and Hispanic students to have worse suspension rates given the evidence of racial disproportionality in school discipline (see for instance Barrett, McEachin, Mills, & Valant, 2017). Finally, rural and town-based schools seem to have lower suspension rates than schools in suburban and urban settings, and smaller schools have lower suspension rates than larger schools. In contrast, suspension rate change measures are very weakly correlated with all school characteristics and student demographics, with no correlation greater than .05. Given the recent push to reduce suspensions in many districts around the state, it may be that this a time of flux for a measure of suspension rate change.

EL progress status and change measures. The EL progress status measure is consistently higher in schools with more White and Asian students (correlations in the .20 to .30 range) and consistently lower in schools with more Hispanic, socioeconomically disadvantaged, and EL students. The latter correlation is the strongest at -.38—EL performance levels are especially low in schools where the proportion of EL students is higher. The association with school characteristics is generally weak, except for the correlation with the log of enrollment, which is .25. This suggests that EL student status is better in larger schools—this may imply that these schools have more resources to serve EL students. EL progress change measures are weakly related to all demographic and school characteristics ($r < .05$ for all characteristics).

Summary of results. Overall, the fairness analyses largely conform with prior literature. Status-based measures, especially of test scores, are moderately to highly correlated with student demographic characteristics. Suspension rates are also moderately correlated with certain demographic variables and seem to vary somewhat across urbanicity types. EL progress varies, especially with the Hispanic and EL concentrations in their schools. All of the change measures, in contrast, are weakly related with all of the demographic and school characteristics variables. A next step would be deeper analyses of these data to tease out whether the observed associations are due to bias or due to do differential effectiveness of schools according to these characteristics.[12]

Consequences

While the Dashboard system is new, we can begin to analyze the likelihood that it will achieve its intended consequences by considering the responses of California superintendents to a series of questions about their beliefs about the purpose of the Dashboard, their efforts to

[12] It would also be interesting to consider how much of the variation in each measure is explained by the full set of demographic and other characteristics that are outside schools' control.

use the Dashboard, and their assessment of the pressure to improve on the Dashboard. We also rely on the response of California voters on their views of the Dashboard.

Evidence from California superintendents. California superintendents largely think the purpose of the Dashboard falls into one of two categories: 34 of 91 respondents said it was to inform the community, and 28 of 91 respondents said it was to provide a progress report on improvement efforts (respondents were allowed to provide multiple answers). Much smaller numbers said the main purpose was to inform parents (13), replace the previous API accountability measure (9), or be used by realtors to describe school quality (2). Superintendents' responses to this question are consistent with a belief that the main purpose of the Dashboard is to align with and provide information about their efforts to achieve their Local Control Accountability Plan. This is largely in line with policymakers' intentions, though the lack of attention to parents needs may be of concern.

The Dashboard is intended to provide users with a holistic view of school performance across multiple measures and for multiple subgroups. We thus explore the extent to which superintendents are using the full breadth of the information provided. Superintendents' interview responses suggest that they are using a very narrow slice of the Dashboard at present. Of the 91 respondents, 39 said they had used the academic performance measures and 36 said they had used the discipline measures. Much smaller numbers (6-17, depending on the measure) said they had used the graduation, EL progress, or college/career readiness measures. When asked which measures they had used when learning about struggling schools, again they overwhelmingly indicated the academic performance measures (39) as opposed to the graduation, discipline, or EL progress measures (5-17). Clearly, the system will not have its intended effects if educators are using a narrow slice of the data. There may be a need for more training or other supports. We do note here that, though the state believes the Dashboard should useful to local decisionmakers, once-yearly accountability measures are unlikely to ever be useful in that context. Furthermore, lack of utility to local decisionmakers does not mean the system might not improve student outcomes—NCLB measures were not useful to local educators, for instance.

Finally, the Dashboard is of course intended to create pressure and incentives for school districts to improve. Superintendents were asked whether they felt pressure and what the source of that pressure was. The large majority (61) said they did feel pressure to improve. Of these, most said the pressure was intrinsic or internal (40) or coming from the community (20). In contrast, very few said they felt pressure from the state (9), the county (3), or the media (3). Again, these responses are quite consistent with the stated theory of action for these reforms, which rely on internal pressure and community pressure but not pressure from counties or the state. Whether these pressures will be effective in generating improvement is, of course, an open question.

Evidence from California voters. Relatively few California voters have heard of the Dashboard, and even fewer have been to the Dashboard website. Among all voters, 16% report having heard about the Dashboard and knowing something about it, 29% report having heard of the Dashboard but not knowing much, and 55% report having not heard of it at all. Parents are substantially more likely to have heard of the Dashboard than the general public. Among parents, 66% have heard of the Dashboard (38% knowing something about it); among other voters, just 38% have heard of the Dashboard (9% knowing something about it). These numbers must increase if the Dashboard is to have its desired effects (though the critical mass of voters that needs to know is an interesting empirical question). The state should consider innovative publicity and community engagement strategies that get beyond merely posting information on a website.

The proportion of Californians who report having been to the Dashboard website is vanishingly small. Just 27% of California voters who have heard of the Dashboard report having been to the website. This means that just 12% of California voters claim to have been to the website. Among parents, 35% have been to the website, versus just 5% of non-parents. There are also troubling gaps in these visiting statistics based on education level. For example, 16% of voters with a college degree or more have been to the website, versus just 8% of those with less than a college degree. Overall, the Dashboard website is not particularly visible to the general community, or even to parents. This is an impediment to the intention of providing valuable information that parents can use to understand school quality and push for change or better choices for their children.

And yet, California voters are enthusiastic about the Dashboard at this point. Among those who have heard of the Dashboard, 51% have a positive impression versus just 12% who have a negative impression. Parents who've visited the Dashboard are even more enthusiastic—72% of them have a positive impression versus just 9% who have a negative impression. When we showed respondents images of the Dashboard, this support remained high—61% said they had a positive impression versus 23% negative. And for parents, it was 76% positive, 17% negative. These baseline numbers suggest that the consequences of the Dashboard are in line with California voters' interests and expectations. Whether this will remain true after more people become familiar with the system is an important question.

While the Dashboard attempts to provide a more holistic view of school progress to parents and the general public, it is not necessarily true that they would prefer this view. In fact, some California voters continue to express interest in having one summative measure of school performance to rely on. We asked this question two ways, randomly assigning the sample to ensure responses could be compared (again, because we did not know the "right" way to ask this question). When simply told that some states provide an overall rating and others do not, about one third of Californians (35%) said they would prefer an overall rating versus 58% who said schools should receive individual ratings on multiple measures (as the Dashboard does). When informed that more that 40 states provide an overall rating, the support was almost exactly split—47% in favor of a single rating, 44% for multiple ratings. However, regardless of how the question was asked, parents (who have the most at stake in the Dashboard) supported a single overall rating by a small majority. With the first wording, support for the single measure was 53/42. With the second wording, support was 57/37. These findings suggest that many voters, and a majority of parents, would like to see a single rating. They may turn south on the Dashboard if the state or an outside entity does not provide one.

Discussion and Implications

The California Dashboard is one year old. Critics have accused it of "setting low expectations" by failing to meaningfully hold schools accountable for subgroup performance (Fensterwald, 2017), and questioning its complexity and the arbitrariness of its cut scores (Bellwether Education Partners, 2017). And, to be sure, there are technical challenges with the system, some of which we have mentioned above. We add to these critiques by providing an initial evaluation of the Dashboard after one year of implementation. Overall, the Dashboard appears to align well with what we infer to be the state's theory of action and seems to have broad support from parents and voters around the state. Our analysis of the Dashboard data, superintendent interviews, and parent poll data has revealed several strengths and several areas for possible improvement.

Strengths

Public support for the Dashboard is high among parents and all California demographic groups at this early stage of implementation. This is likely because, despite some voters who wish to see only one metric of school progress, the Dashboard aligns with the views of most voters that school effectiveness is defined by much more than just mathematics and ELA test scores. The early evidence on the construct validity of the system is middling—the measures are broader than under NCLB, they generally correlate in the expected directions, and the use of change measures gets the state much closer to rating schools on their contributions to student outcomes as opposed to just who they happen to enroll. But the state's lack of a clear rationale and guiding goal for the Dashboard makes it hard to say whether it is well designed for that purpose, and it still largely measures schools on who they enroll rather than how effectively they educate those students. Using change measures also improves the fairness of the system, reducing to nearly zero the relationship of those ratings with student demographics (though the relationships of status measures to student demographics are still quite strong). The subgroup size chosen by the state dramatically increases the number of student subgroups that are brought into the light of day as compared to the old system. Voters describe the Dashboard as highly transparent after being presented images of it. These are promising signs that the Dashboard is starting from a place of enthusiasm.

Areas for Improvement

Improving awareness and use. Though enthusiasm for the Dashboard is high, awareness is very low. A bare majority of parents report being aware of the Dashboard, and an even smaller proportion of non-parents are. Few parents report having used the Dashboard website, and almost no non-parents have. More educated voters are twice as likely to report having used the website as less educated parents. In short, awareness and use of the Dashboard is very low and must improve. We suggest that the state invest in publicizing the

Dashboard, especially in communities that could benefit from it the most (low-income communities where the LCAP process is likely more consequential).

Superintendents also report problems with the utility of the Dashboard. The major critique was that the data in the Dashboard were incomplete and out-of-date relative to other data sources they had. If the Dashboard data really are to inform the LCAP process and be of use to district leaders, the data must be made available rapidly and must be aligned with what superintendents care about. Or perhaps it is the case that no single system can serve both purposes, and the state needs to target the Dashboard at parents and some other system at local educators. More than a quarter of the interviewed superintendents did not report using the Dashboard data for any purpose. While this number may go down over time, it suggests that, for many California superintendents, the Dashboard is providing no added value.

Improving comparability. Analysis of the Dashboard data indicates that the status measures of performance that make up half the Dashboard ratings are highly related to school and student characteristics. While this is to be expected, the state could do more to make the system fairer to diverse schools. One popular solution, which many superintendents seem to support, would be to at least allow users to make comparisons that adjust for local context. For example, while the state could keep the front page of the Dashboard the same, they could give users the ability to compare schools with similar demographics or in similar locations. Superintendents clearly stated this as a preference. And while we did not ask voters this question directly, voters and parents would undoubtedly appreciate the ability to more directly compare schools for purposes of making decisions and informing judgments. The Dashboard offers little opportunity to do so.

Providing better information about schools' contributions to student learning. The state has made some strong choices in the selection of measures to include in its accountability system. For instance, the use of average scale scores rather than percent proficient aligns much better with what is known about status-based measures of performance (Polikoff, 2016). However, the state has chosen a simplistic "change" measure by merely taking the difference between this year's scores and last year's scores on each outcome. This approach suffers from many problems, not the least of which is that it does not adjust for the fact that these are different students being compared to one another (i.e., there are "cohort effects"). Especially for test scores, where there is a wealth of knowledge about the best ways to construct accountability system growth measures, there is no reason for the state to choose the approach it did. The state should choose a more appropriate growth measure, such as a two-step value-added model (Ehlert, Koedel, Parsons, & Podgursky, 2014).

Improving subgroup transparency. One of the main criticisms of the Dashboard, especially from civil rights groups, is that the Dashboard obfuscates the performance of subgroups (Fensterwald, 2017). There is some truth to this criticism—the front page of the Dashboard, the so-called "equity report," reports overall performance, but details on subgroup performance are mostly relegated to other pages in the report. And schools can earn strong overall ratings even if subgroup performance is poor. When combined with the fact that there are limited consequences for poor performance under the Dashboard (and, as our superintendents indicated, there is very little pressure from the state to improve), the Dashboard may not be a strong tool for promoting equity in its current form.

There are several possible approaches that could remedy this situation. First, the state should consider lowering subgroup size further. Lowering from 30 to 20 would dramatically increase the proportion of students whose performance would count towards subgroup reporting. Many states have subgroup sizes 20 or smaller; California should follow suit. Second, the state should explore ways to make subgroup results more visible on the Dashboard report. For instance, one approach would be to allow users of the website to indicate a subgroup of interest and have the website then foreground that subgroup on each school's report. Regardless of the approach taken, the state must ensure that the Dashboard website helps schools address the performance gaps that are at the heart of LCAP, and the current version falls short in that regard.

Conclusion

The California Dashboard is a bold experiment at bringing the "California Way" to school accountability. The state is bucking the trend of other states, avoiding giving schools a summative rating and making it all-but-impossible to directly compare schools' performance. Presumably, the intention is that competition and ratings—the hallmark of NCLB and previous accountability systems—drive unproductive behaviors. Thus, the new system moves sharply away from rankings, grades, and punishments.

Whether the experiment will work is an important question for future research. For now, there are promising signs but also areas where the state could shore up the dashboard with little effort. The fundamental premise of the CA Dashboard accountability system is that when given the right information and adequate support, schools can use it to improve. We suggest the same is true of the new "California Way" approach to accountability: while there are many things the Dashboard is doing well, we hope that the state can use the insights discussed in this report as a guide as they improve the system in the upcoming years.

References

Ahn, T. & Vigdor, J. (2014). *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina (NBER Working Paper No. 20511*). Cambridge, MA: National Bureau of Economic Research.

Barrett, N., McEachin, A., Mills, J. N., & Valant, J. (2017). *What are the sources of school discipline disparities by student race and family income?* New Orleans, LA: Education Research Alliance for New Orleans.

Bellwether Education Partners. (2017). *An independent review of ESSA state plans: California.* Washington, DC: Author.

Bonilla, S., & Dee, T. (2017). T*he effects of school reform under NCLB waivers: Evidence from focus schools in Kentucky (NBER Working Paper No. 23462)*. Cambridge, MA: National Bureau of Economic Research.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268.

California Department of Education. (2017b). *California School Dashboard—Core messages.* Sacramento, CA: Author. Retrieved from https://www.cde.ca.gov/ta/ac/cm/dashboardcoremessages.asp.

California Department of Education. (2017a). *California School Dashboard: Technical guide 2017-18 academic year.* Sacramento, CA: Author.

Chakrabarti, R. (2014). Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics, 110*, 124–146.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, 12(3), 311–329.

Davidson, E., Reback, R., Rockoff, J. E., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. *Educational Researcher, 44*(6), 347-358

Dee, T. S., & Jacob, B. A. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*, 418–446.

Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis, 35*(2), 252-279.

Deming, D., Cohodes, S., Jennings, J., & Jencks, S. (2013). *School accountability, postsecondary attainment and earnings (NBER Working Paper No. 19444).* Cambridge, MA: National Bureau of Economic Research.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). Choosing the right growth measure. *Education Next, 14*(2), 66-71.

Fensterwald, J. (2017, March 15). California's long-awaited School Dashboard debuts. *EdSource*. Retrieved from https://edsource.org/2017/californias-long-awaited-school-dashboard-debuts/578687.

Figlio, D. N., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. J. Machin, & L. Woessmann (Eds.), *Handbooks in economics: Economics of education* (Vol. 3, pp. 383–421). NorthHolland, the Netherlands: Elsevier.

Figlio, D. & Rouse, C.E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*(1–2), 239–255.

Figlio, D.N., Rouse, C.E., & Schlosser, A. (2009). *Leaving No Child Behind: Two paths to school accountability*. Working Paper.

Gaddis, S. M., & Lauen, D. L. (2014). School accountability and the black-white test score gap. *Social Science Research, 44*, 15-31.

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis, 36*(4), 417-436

Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic staffing? How performance pressures affect the distribution of teachers within schools and resulting student achievement. *American Educational Research Journal*, *54*(6), 1079-1116.

Hamilton, L. S., Berends, M., & Stecher, B. M. (2005). *Teachers' responses to standards-based accountability*. Santa Monica, CA: RAND.

Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments (RAND Working Paper WR-374-EDU)*. Santa Monica, CA: RAND.

Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal, 38*(2), 289-318.

Hough, H., & Miller, R. (2017). *Letter to the California State Board of Education.* Retrieved from http://coredistricts.org/wp-content/uploads/2018/05/reformattedCORE-PACE_May17_SBELetter_20170504.pdf.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89(5–6)*, 761–796.Johnson, R., & Tanner, S. (2018). *Money and freedom: The impact of California's school finance reform*. Palo Alto, CA: Learning Policy Institute.

Jacob, B. A. & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843–877

Jennings, J. L. & Bearak, J. M. (2014). ''Teaching to the test'' in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher,* 43(8), 381-389.

Jennings, J., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education, 87*(2), 125-141.

Johnson, R. C., & Tanner, S. (2018). *Money and freedom: The impact of California's school finance reform.* Palo Alto, CA: Learning Policy Institute.

Judson, E. (2013). The relationship between time allocated for science in elementary schools and state accountability policies. *Science Education, 97*(4), 621-636.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic perspectives*, *16*(4), 91-114.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 9*(49).

Lauen, D. L. & Gaddis, S. M. (2012). Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability pressure on student achievement. *Educational Evaluation and Policy Analysis 34*(2), 185–208.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, *32*(7), 3–13.

Loeb, S. & Strunk, K. (2007). Accountability and local control: Response to incentives with and without authority over resource allocation and generation. *Education Finance and Policy, 2*(1), 10–39.

Markowitz, A. J. (2018). Changes in school engagement as a function of No Child Left Behind: A comparative interrupted time series analysis. *American Educational Research Journal.* Published online before print.

McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

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

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, *92*, 263–283.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Polikoff, M. S. (2012a). Instructional alignment under No Child Left Behind. *American Journal of Education,* 118(3), 341–368.

Polikoff, M. S. (2012b). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis*, 34(3), 278–294.

Polikoff, M. S. (2013). Teacher education, experience, and the practice of aligned instruction. *Journal of Teacher Education,* 64(3), 212–225.

Polikoff, M. (2016, July 12). A letter to the U.S. Department of Education (final signatory list). *On Education Research*. Retrieved from https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education/.

Polikoff, M. (2018). The PACE/USC Rossier California education poll. *CEPEG Blog.* Los Angeles, CA: Center on Education Policy, Equity and Governance. Retrieved from https://cepeg.usc.edu/pace-usc-rossier-california-education-poll/.

Polikoff, M. S., & Korn, S. (in press). School accountability. In J. G. Dwyer (Ed.) *Oxford handbook of children and the law.* Oxford, UK: Oxford University Press.

Polikoff, M. S., McEachin, A., Wrabel, S. L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher, 43*(1), 45-54.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy, 6*(3), 207–241

Reckase, M. (1997). *Consequential validity from the test developer's perspective.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sass, T. (2008). *The stability of value-added models of teacher quality and implications for teacher compensation policy.* Washington, DC: CALDER/Urban Institute.

Spillane, J. P. (2004). *Standards deviation: How schools misunderstand education policy*. Cambridge, MA: Harvard University Press.

Sun, M., Saultz, A., & Ye, Y. (2017). Federal policy and the teacher labor market: exploring the effects of NCLB school accountability on teacher turnover. *School Effectiveness and School Improvement*, 28(1), 102–122.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. J*ournal of Research on Educational Effectiveness*, 8(2), 245–279.

Table 1
Correlations of Dashboard Measures for Elementary/Middle Schools

	Math achievement levels	ELA achievement levels	Suspension levels	EL progress levels	Math achievement change	ELA achievement change	Suspension change	EL progress change
Math achievement levels	1.00							
ELA achievement levels	0.94	1.00						
Suspension levels	-0.45	-0.33	1.00					
EL progress levels	0.30	0.44	0.15	1.00				
Math achievement change	0.22	0.16	-0.13	0.00	1.00			
ELA achievement change	0.09	0.14	-0.04	0.05	0.64	1.00		
Suspension change	0.01	-0.01	0.26	0.01	-0.09	-0.08	1.00	
EL progress change	-0.01	0.03	-0.09	0.36	0.02	0.05	-0.02	1.00

Table 2

Correlations of Dashboard Measures for High Schools

	Math ach. levels	ELA ach. levels	College/career readiness levels	Suspension levels	Graduation levels	EL progress levels	Math ach. change	ELA ach. change	Suspension change	Graduation change	EL progress change
Math ach. levels	1.00										
ELA ach. levels	0.91	1.00									
College/career readiness levels	0.81	0.77	1.00								
Suspension levels	-0.30	-0.31	-0.23	1.00							
Graduation levels	0.54	0.57	0.58	0.05	1.00						
EL progress levels	0.25	0.28	0.23	-0.04	0.28	1.00					
Math ach. change	0.27	0.19	0.05	-0.03	0.06	0.05	1.00				
ELA ach. change	0.19	0.29	0.06	0.01	0.11	0.01	0.65	1.00			
Suspension change	0.03	0.03	-0.01	0.22	0.03	0.01	0.22	0.02	1.00		
Graduation change	-0.19	-0.17	-0.18	-0.07	0.04	0.03	0.03	0.05	-0.03	1.00	
EL progress change	0.00	0.01	-0.01	-0.03	0.01	0.51	0.03	0.03	0.00	0.06	1.00

Table 3

Proportion of Schools Reporting Each Subgroup Under Various Scenarios

	Previous subgroup sizes (100)	Dashboard subgroup size (30)	Proposed subgroup size (20)
African American	14%	34%	43%
American Indian/Alaska Native	0%	1%	3%
Asian	18%	38%	46%
Filipino	4%	15%	23%
Hispanic	85%	93%	95%
Pacific Islander	0%	1%	3%
White	53%	71%	78%
Multiple Races/Two or More	3%	29%	40%
English Learner	62%	82%	86%
Socioeconomically Disadvantaged	89%	95%	96%
Students with Disabilities	36%	84%	90%
Average total number of subgroups	3.65	5.44	6.04

Table 4

Proportion of Students Included in Subgroup Accountability Under Various Subgroup Size Rules

	Previous subgroup sizes (100)	Dashboard subgroup size (30)	Proposed subgroup size (20)
African American	38.8%	13.0%	8.3%
American Indian/Alaska Native	94.4%	79.0%	71.3%
Asian	23.9%	7.8%	4.7%
Filipino	64.3%	31.3%	20.6%
Hispanic	1.3%	0.2%	0.1%
Pacific Islander	98.9%	85.0%	72.0%
White	7.7%	2.1%	1.1%
Multiple Races/Two or More	79.7%	27.4%	15.7%
English Learner	8.4%	1.3%	0.6%
Socioeconomically Disadvantaged	0.9%	0.2%	0.1%
Students with Disabilities	38.1%	2.7%	1.0%