Abstract

School accountability systems in the United States have been criticized on a number of fronts, chiefly on grounds of completeness and fairness. This study examines an alternative school quality framework—one that seemingly responds to several core critiques of present accountability systems. Examining results from a pilot study in a diverse urban district, we find that this alternative system captures domains of school quality that are not reflected in the current state system, specifically those measuring opportunity to learn and socioemotional factors.Further, we find a less deterministic relationship between school quality and poverty under the alternative system. We explore the policy implications of these findings vis-à-vis the future of accountability.

Keywords: Accountability, Equity, School Quality

Imagining the Future of Accountability:

Pilot-Test Results from a Holistic School Quality Measurement System

Over the past two decades, policy leaders have establishededucational measurement and accountability systems in all 50 states. These systems are intended to help policymakers identify schools in need of support and intervention, to inform and empower the public, and to establish clear and consistent goals for educators and school leaders.

Whatever their successes, however, these systems continue to face a number of challenges, particularly with regard to capturing the multifaceted nature of school quality. Schools serve many purposes and advance multiple aims through a variety of interconnected practices (Author, 2017; Figlio& Loeb, 2011; Ladd & Loeb, 2013; Rothstein & Jacobsen, 2006). For political and practical reasons, however, current data systems have generally focused on a set of fairly basic outputs, chief among them being student standardized test scores in math and English.As a result, these measurement and accountability systemshave been roundly criticized—forfailing to capture the full picture of school quality, for relying too heavily on measures linked to student demography, and forproducing a range of unintended consequenceslike curricular narrowing and teaching-to-the-test (e.g. ASCD, 2014; Cowley, 2006; Darling-Hammond, 2004; NEA, 2011, Spalding, 2014).

Our study examinesthepilot year of a holistic school quality measurement system built for a mid-sized, highly diverse urban district in Massachusetts. Constructed with stakeholder input, and responding to severalcommon criticisms of presentefforts to measure school quality, this system represents a model of what accountability systems of the future might look like. Insofar as that is the case, analysis of it may help answer some of the key questions policymakers face as

they revisit state-level measurement and accountability systems under the Every Student Succeeds Act (ESSA).

Ultimately, our analyses reveal thatalthough there is some correlation between the current state accountabilityratings and the ratings generated by a more comprehensive set of data, the inclusion of additional measures can dramatically alter the overall interpretation of school quality. As we find, measurement and accountability systems that draw chiefly on achievement scores in math and English incompletely capture school performance, strongly reflect demographic variables, and consequently may foster the mistaken view that school quality is a uniform concept. In short, although a more holistic approach to measurement does not represent a perfect solution, it does appear to offer a much more viable basis for future accountability systems.

Literature Review

Current DataCollection and Reporting Practices

No Child Left Behind (NCLB)dramatically expanded the amount of information collected by states about school and district performance, using student outcomes to construct accountability systems with meaningful consequencesSpecifically, the law required states to collect and report data on achievement and teacher quality, and to disaggregate data by student subgroups (U.S. Department of Education, 2013). Over the next decade, schools and districts bristled under NCLB, and scholars highlighted a broad range of flaws in the law.Ultimately, theU.S. Department of Education (USED)responded by initiating awaiverprocess that led to significant changes in accountability in many states, such as the elimination of requirements surrounding Adequate Yearly Progressand the goal of 100 percent proficiency.Although waivers were granted to many states, they largely served to limit and structure data collection and

reporting, rather than to expand or reimagine those enterprises. Consequently, despite otherwise significant variance across states with regard to education, data collection and reporting practices continued to display striking similarity. AsMikulecky and Christie reported in 2014, all states incorporated student achievement and graduation rates into school accountability systems, with a majority including student growth, gap closure, and proxies for postsecondary and career readiness. This largely remains the case under ESSA, which replaced NCLB in late 2015, thoughstates continue to revise their accountability frameworksin preparation for the 2017-2018 school year when the law will go into full effect.

State measurement and accountability systems do, of course, include more than just test scores in math and English. Still,these systems largely fail to address the full range of what schools do. Several core subjects, for instance, go untested, and therefore unmeasured. Current measurement and accountability systems largely ignore aspects of student physical, social, and emotional health emphasized by schools (Author, 2017; Downey, von Hippel, & Hughes, 2008; Mintrop&Sunderman, 2009). And,these systems largely fail to provide information that might lead to meaningful improvements in curriculum, teacher preparation, and school resources (Darling-Hammond, 2007). Moreover, research suggests that various elements of school quality are not intrinsically aligned, indicating that a measurement system designed to capture some elements of school quality will not necessarily capture others (e.g. Rumberger & Palardy, 2005).

Current measurement and accountability systems have also been criticized for measuring factors that are not largely under the control of schools. As research has revealed, student standardized test scores tend to correlate strongly with student demographic characteristics (Reardon, 2011; Davis-Kean, 2005), and school rankings tend to correlate strongly with school-level poverty (Spalding, 2014). Consequently, test-based outcome variables may tell

stakeholders less about school performance than about families and neighborhoods.Insofar as that is the case, such dataoffer little in the way of actionable information, and may unnecessarily stigmatize schools with diverse student bodies.

Finally, because measurement and accountability systems shape school and district goals and activities, researchershave also explored the degree to which narrow conceptions of school quality have produced troubling unintended consequences. As scholars have shown, current systems have led to narrowing of the curriculum and an increased emphasis on test preparation (Dee, Jacob & Schwartz, 2012; Hamilton, et al., 2007; Jennings &Bearak, 2014). Additionally, such systems have fostered an environment in which teachers are less satisfied with their jobs (Markow&Pieters, 2012), and in which students exhibit higher levels of stress (Segool et al., 2013).

Measurement and accountability systemsremain in place as cornerstones of governance, and ESSA, like NCLB before it,continues to place significant weight on achievement scores and graduation rates in school performance rankings. That said, there are twoprimary reasons to suspect that data systems will change under the new law. First, ESSA stipulates that states must use an additional metric in tracking student success—a measuresuch as student engagement, school climate, or access to advanced coursework. Second, the additional flexibility in ESSA, both real and perceived, is likely to spur reforms in areas that have been unpopular with educators or that have triggered scholarly criticism. Influential groups such as ASCD and the National Education Association (NEA) have strongly advocated for the inclusion of multiple measures in determinations of school success (ASCD, 2014; NEA, 2011), as have prominent academics (e.g. Darling-Hammond et. al, 2016).

Emerging Efforts to Measure School Quality More Comprehensively

What other dimensions of school quality, beyond achievement results in math and English, might be measured? And what is the relationship between those new measures and the test scores that presently dominate state accountability systems?

Much discussion has revolved aroundOpportunity to Learn (OTL) measures, which are presumably less closely tied to student demography and more informativeabout what is going on inside schools. Model OTL frameworks, like that of the National Council for Teacher Education, tend to emphasize school culture, teaching environment, learning resources, and resources from the community (NCTE, 2012).

There has also been a great deal of discussion about expanding measurement to include Social and Emotional Learning (SEL). For the past two decades, groups like the Collaborative for Academic, Social, and Emotional Learning (CASEL) have been advocating for a greater emphasis on so SEL, as have organizations like the National Association of State Boards of Education (NASBE, 2013). Presently, three states—Illinois, Kansas, and Pennsylvania—have adopted comprehensive SEL standards with developmental benchmarks.

Research has demonstrated a connection between OTLand SEL variables on the one hand, and student standardized test scores on the other. Shindler, et al. (2009), for instance, found correlations of 0.5 to 0.7 between measures of school climate and student achievement in California schools. Similarly, in a study of Chicago schools, Erbe (2000) found that "focus on learning," "school commitment," and "parental involvement" had roughly 0.5 to 0.7 correlations with math achievement. Some studies have also shown that specific school resources (e.g. funds used to support targeted instruction) have an effect on achievement (Archibald, 2006; Lavy, 2012), though such findings are not universally true across the scholarly literature (e.g. Hanushek, 1997; Hanushek, 2003; Houtenville, & Conway, 2008). Numerous other studies

document the connections between various measures of school quality and outcome measures of student achievement (Berkowitz, Moore, Astor, &Benbenishty, 2016). Cadima, Peixoto, & Leal, 2014; Darling-Hammond, 2000; Kutsyuruba, Klinger, & Hussain, 2015; Lubienski, Lubienski, & Crane, 2008).

Still, it seems unwise to validate various measures of school quality solely by establishing relationships with standardized achievement scores. Perhaps the most compelling argument against such a practice is the fact that different domains of school quality may be orthogonal, which is to say that successes in some areas may not coincide with successes in others. The most relevant research in this area comes from investigations into teacher effectiveness. Jackson (2016), for instance, finds that teachers exhibit variability in their effect on student behaviors, including suspensions, attendance, course grades, and on-time grade progression, as well as longer-term outcomes such as high school completion. Furthermore, these teacher effects on non-test score outcomes exhibit only weak positive correlation (ρ=0.16) with a teacher's value-added to standardized achievement. A comparable study, which uses data from over 1 million students in the Los Angeles Unified School District, reaches similar conclusions about the multidimensionality of teachers (Petek& Pope, 2016).In other words, teachers can be relatively strong in raising student achievement without improving student behavioral outcomes, and vice versa. These and other studies (e.g. Grissom, Loeb, & Doss, 2015) provide compelling evidence that teacher effectiveness is not a unidimensional construct, which in turn suggests that school quality is not either.

A number of districts are currently employing school quality frameworks that extend beyond academic achievement by including measures of OTL and/or SEL. The Chicago Public Schools for instance, have worked with the University of Chicago's Consortium on School

Research to employ the 5Essentials framework. This framework measures the effectiveness of school leaders toimplement a clear and strategic vision, the level of support for teachers, the involvement of families, the safety and orderliness of the school, and the level of academic challenge in classes. Recently, the entire state of Illinois adopted the framework, making the case that"test scores alone do not provide a full picture of teaching and learning in any one school" (5Essentials, n.d.). The 5Essentials survey is taken by all pre-K through 12th grade teachers, as well as by all 6th-12th grade students in Illinois, and reports generated from these surveys are produced for all schools in the state.

Similar work is currently being done by the California Office to Reform Education (CORE)—a consortium of districts that collectively educate over one million students in the state. CORE's School Quality Improvement Index is built around a 100-point scale: 60 points allotted for the academic domain and 40 for social-emotional and school culture factors. Within the academic domain, two-thirds of points are determined by test scores, with raw scores and growth scores counting equally. The remaining third of the academic domain is determined by graduation rates. For the 40 points allotted to social-emotional and school culture factors, the CORE districts rely on a broader range of measures, including how many students are missing significant amounts of school, how many are suspended or expelled, and how many English Language Learners have become fluent. Additionally, the CORE districts plan to incorporate results fromschool climate surveys given to students, parents, and teachers—a practice that is increasingly supported by research (e.g. Kane &Staiger, 2012; Wilkerson, et al, 2000).

Given criticism from scholars and the public,as well as new flexibility afforded by ESSA legislation, it appears that state-level measurement andaccountability systems will expand in coming years. As they do, it seems likely that they will includemany of the input measures that

fall under the umbrella of OTL, and also many of the outcome measures included in SEL frameworks.

Current Study

Study Population

This project took place in a diverse, mid-sized urban district in the state ofMassachusetts. In the 2014-2015 school year, nearly half of students in the district were Hispanic, roughly a third were white, and the remaining students were mostly Asian and African American (MDESE, 2016). The district had nearly 5,000 students spread out across its one early learning center, seven traditional primary schools, two alternative schools, and one secondary school. The district serves a fairly high-need population, withover one-third of students being deemed economically disadvantaged and nearly one in five being categorized as an English Language Learner.

The "Beyond Test Scores" Project

In the spring of 2014, our research team partnered with the district, which was interested in measuring school quality "beyond test scores." District administrators, the city's school committee, and other civic leaders had expressed frustration with a narrow range of measures, which they argued had limited their ability to track progress across many aims, and which appeared tocorrelate strongly with student socioeconomic status.A number of stakeholder groups expressed a particular interest in expanding the current conception of school quality to include OTL and SEL measures.

Our team began by compiling an inventory of school quality factors distilled from national polling, educational research, and community surveys. This generated a list of several dozen potential variables. Some of these variables repeated each other, differing primarily in the

language used to express them. In those cases, we simply selected the factor with the clearest wording. Additionally, many factors on the list seemed to be of different grain size—some, for instance, were quite specific, while others seemed to be umbrella concepts for multiple factors. In those cases, we retained the smaller, specific items, and set aside the umbrella terminology for later in the process.

Having distilled 32 separate factors for a school quality framework, we organizedthem into a hierarchical taxonomy. In doing so, we paired together similar metrics, like "student sense of belonging" and "student-teacher relationships," into measures—in this case: "Relationships." Next, we nested our 16 subcategories under five major categories. The "Relationships" measure, for instance, together with "Safety" and "Academic Orientation," formed a major category: "School Culture." This approach allowed us to preserve a high level of complexity, while also respecting the limits of working memory (Baddeley, 1992; Cowan, 2001).

Throughout this process, we conducted focus groups with stakeholders in the community. Ultimately, we conducted ten focus groups—three with teachers, two with principals and administrators, and five with parents and community members. Though educators, administrators, and laypeople have different priorities and concerns, these different constituencies were able to agree on a single framework. As one of our research assistants concluded in a memo analyzing results from focus groups with principals and community members: "there was virtually no disagreement between the two groups." We found similarly strong overlap with results from our focus groups with teachers. No longer hearing new suggestions from our stakeholders—a point of "saturation" (Morse, et al., 2002) in our sampling—and seeing no major disagreements among them, we sent a copy of the framework to district leaders for their approvaland adoption. Note that the information gathered in these focus

groups was used to engage and inform stakeholders in the construction and validation of the alternative framework; they did not serve as a collection point for the school quality data that is analyzed in this study. See Appendix A for a copy of the School Quality Framework (SQF), and see Author (2017) for much more detail about the creation of the SQF and the use of focus groups in that process.

Our purpose in this articleis not to make claims about the effectiveness or generalizability of this framework. Instead, we have provided an overview ofits design in order to show that it is an adequate test case for the state-level measurement and accountability systems that will emerge in coming years. It includes many of the most commonly discussed OTL measures, as well as a number of SEL measures. Additionally, its development incorporated the feedback and multiple stakeholder groups, and to a large degree addresses the public concerns that led to changes in federal law pertaining to school accountability.

Data and Methods

There are three primary sources of data used in this study: a survey of students in the district,a survey of teachers in the district, and a collection of administrative data made available by the district. The unit of analysis in this study is the school. The sample includes seven elementary/middle schools, five of which are pre-K-8, one of which is K-8, and one of which is K-6. To ensure comparability, the only high school in the district was excludedfrom analysis, as were an early childhood center and two small alternative schools. These excluded schools serve unique populations, andtherefore do not lend themselves to norm-based comparison with the included schools.

Students in grades 4 and above at each of the elementary/middle schools were issued perception surveys, with students in grades 3 and below being excluded due concerns of age

appropriateness, specifically with regard to reading comprehension level. The survey produced a student sample of 1607 students, or 98.2 percent of non-excluded population. The teacher survey was issued to and completed by all 229 teachers within the sample schools. Administrative data were collected at the end of the academic yearfor all schools in the sample.

Surveys were constructed using established scales when available. The internal consistency of all survey scales was examined, and Cronbach's alpha was calculated for all scales. In order to examine the relationship between data sources, all metrics were normalized to have a mean of zero and a standard deviation of 1.0. Thus, the analytic approach used in this study includes district-normed measures. Consequently, school quality is calculated in comparison to other schools within the district analytic sample.In other words, a zero-sum approach is taken, where a school's quality is judged with respect to how the other schools in the district perform within a given metric.[1]

We examine the correlations of metrics within the five major categories of school quality in our framework. Note that all correlations presented in this articleare at the school level, which aligns with our interest in understanding how factors of school quality relate. We then sum all metrics within a given school measure to form that school quality measure—first at the submeasure level, then the measure level, and finally at the major category level.

Next, we examine the correlations between major school quality categories. We generally expect to see positive correlation, as schools that perform well in one domain might be expected to succeed in others as well. However, we do not anticipate strong correlations, as a general rule, as we also expect that schools will exhibit relative strengths and weaknesses. Again, there are theoretical reasons to suspect that different measures of school quality are orthogonal to various degrees, so one would expect commensurate empirical variability as well.

We then form a composite (i.e. overall) SQF ranking by summing combined z-scores over all categories. This allows us to examine how school rankings might differ across the alternative and traditional frameworks. We then dive deeper into these relationships, reporting the average school-level correlations between SQF metrics and the state's Progress and Performance Index (PPI), which is currently used by the Department of Elementary and Secondary Education to rate schools, and which is discussed in greater detail later in the article.[2]

Finally, we examine the relationship between each system (the existing state accountability system, and the alternative SQF system) and school poverty. To do so, we calculate the average school-level correlations between the state PPI system andschool poverty, using the percentage of Economically Disadvantaged (ED) students in each school. We then perform the same calculation for the SQF metrics, reporting average correlations for each of the major categories in the SQF toillustrate which domains of the alternative framework mirror school poverty, and which do not. Through these analyses we addresstwo primary research questions:

1. To what extent does a more holistic array of school quality data capture information not reflected incurrent accountability systems? Specifically, whatare the correlationsbetween SQF metrics and the Massachusetts PPI, and how might school rankings differ under the two frameworks?
2. How mightexisting data systems reflectthe out-of-school context tied to student demography, rather than something about school quality? Specifically, whatare the correlations betweenSQF metricsand school poverty, and how do those compare to the correlation between PPI and school poverty?

Findings

Scrutinizing the Alternative Framework

Before conducting the analyses that directly address our research questions, we examined the measures employed in the alternative framework to confirm that they met basic standards. We performed factor loading for the 22 survey-based metrics, calculating the internal consistency of this composite using Cronbach's alpha. We found that 20 metrics exceeded 0.7—long held as a rule of thumb in scale reliability (Nunnaly, 1978)—with the remaining two survey scales exhibiting reliability estimates of 0.69 and 0.58.Two survey-based metrics—Arts Exposure (5Cia) and Physical Activity (5Diia)—were based on a single survey question and therefore did not have associated reliability statistics. We consider these results to be sufficient to conduct the subsequent analyses necessary for this study.

Furthermore, schools exhibited considerable variation on most metrics (see Table 1), as indicated by standard deviations between 0.16 and 1.2 for teacher survey metrics, and 0.09 and 0.46 for studentsurvey metrics, on a five-point Likert scale. The remaining six metrics, which were taken from district administrative records, also displayed meaningful variation in our sample. Given the norm-referenced approach to this study, such variation is a necessary pre-condition for the remaining analyses.

INSERT TABLE 1 ABOUT HERE

In order to illustrate the relationship between the metrics that form school quality categories, we take the within-category average of metrics-level correlations. The average within-category correlations are shown in Table 2, and range from 0.15 (Category 3, Resources) to 0.45 (Category 2, School Culture). These weak to moderate average correlation magnitudes generally support the grouping of such metrics to form school quality categories, as they show a positive, but not deterministic, relationship between these related measures.

INSERT TABLE 2 ABOUT HERE

Wenext analyze the relationships between, as opposed to within, the five major categories of school quality in the SQF. To do so, we aggregate all normalized metrics to form a single categorical school quality score. In an effort to evenly weight categories,we aggregate from lower levels upward.For example, by combining the Class Size ratio metric (3Biia) with the Class Size Scale metric (3Biib), we formed a single submeasure: Class Size (3Bii). Then,we combined Curricular Strength (3Bi) with Class Size (3Bii) to form a single measure: Curricular Resources (3B). Lastly, we combined relevant measures to create a major category score—in this case, combining Facilities and Personnel (3A), Curricular Resources (3B), and Community Support (3C) to form Resources (3). Table 3 shows the Pearson correlation coefficients at the category level. All correlations shown in table 4 are positive, with magnitudes varying from 0.18 to 0.70. Overall, these findings suggest that categories used to construct the framework exhibit meaningful associations, while not being deterministically related.

INSERT TABLE 3 ABOUT HERE

Relationships between the Alternative Framework Metrics and the State System

School Rankings

One common criticism of existing data systems is that they fail to measure many valued aspects of school quality. We now turn to our first research question, initially by exploring how school rankings might align and diverge under the two different frameworks. Specifically, we sought to determine if the holistic picture of school quality is somehow being captured by the current Progress and Performance Index, which is used by the Massachusetts Department of Elementary and Secondary Education to classify schools into one of five accountability and assistance levels. The PPIis a single number between 0 and 100, produced by combiningtest

score results—specifically: progress toward 100 percent proficiency, as well as average test score improvement measured by the state's Student Growth Percentile (SGP)—along with graduation and dropout rates. It is possible, after all, that while the information may not be *presented* in the PPI system, it is neverthless*accounted for*. Table 4 lists the seven study schools in order of their PPI ranking. Presented alongside this ranking scheme are six alternative rankings—ranked, respectively, by each of the five major categories of the alternative SQF model, as well as by the composite, or summative z-score, of those five categories.

Overall, we see some agreement in overall rankings between the two frameworks. Two schools (T and U), for instance, maintain their relative place when comparing the state PPI ranking to the alternative composite SQF ranking. Two schools (W and Y) move one place, while one school (X) moves two places when switching from the PPI model to the SQF model. Two schools moved three spots. School V, which was third according to the PPI, was sixth in the composite SQF model, as it had relatively low scores in all categories except Academic Learning; school Z jumped from seventh place to fourth. It is worth noting here that some categories drive overall SQF rankings much more than others. School Culture, for instance, exhibits a range of 0.85 s.d. across the seven schools. By contrast, the highest and lowest scoring schools in the Character and Wellbeing category are separated by only 0.26 s.dThus, although there is some general alignment between PPI and SQF, differences exist, and may be driven by particular categories.

Perhaps the more important aspect to focus on here, however, is not the overall congruence in rankings between frameworks, but rather the variability across individual measures that speaks to the multidimensionality of school quality. As seen in Table 4, positional changes are more substantial within individual categories than in the composite of all five.

Schools W and X, for instance, are quite similar in terms of PPI rankings—placing in the 53rd and 50th percentiles, respectively. Yet differences abound. School W ranks first in both Teachers and the Teaching Environment and in Character and Wellbeing outcomes; it also ranks second in School Culture. School X, by contrast, ranks last in all three of those categories.

INSERT TABLE 4 ABOUT HERE

Overall, we find that school rankings according to the PPI and SQF models exhibit some approximate alignment, with a few notable exceptions. Perhaps the more important takeaway, however,is that examining only a single index measure—either the PPI or the composite SQF—obscures the fact that some schools perform well in some domains while doing relatively poorly in others. Furthermore, the SQF provides a rich set of indicators which schools might use to help guide policy and improvement efforts; the same cannot be said of the PPI.

SQF Metrics and PPI

We nowexplore our first research question more deeply, reporting oncorrelations between individual metrics from the more comprehensive SQF and the state systemin order to understand what specific information is not being captured by the PPI calculation (see table 5). We find considerable variability not only in the relationships between individual metrics and the PPI, but also between PPI and the average correlations of major categories of the SQF. Moreover, the strengths of the relationships between various SQF metrics and PPI offers suggestive evidence as to the ways in which PPI may inadequately measure a fuller conception of school quality. In general, we find that this relationship is usually stronger when a metric was related to student achievement or family background, and lower when it was more a reflection of educational opportunity.

Metrics within the Teachers and the Teaching Environment category exhibit an average correlation of 0.08 with the PPI. In addition,metrics within this category have very different associations with the PPI. Three metrics within Teachers and the Teaching Environment exhibited moderately negative correlations: teacher perceptions of the usefulness of professional development, student perceptions of the level of teacher interest in students, and principal leadership. In other words, teachers in lower PPI schools exhibited greater interest in students, as measured by student perception surveys, and also found their professional development to be more useful. This serves as a powerful instance of the ways in which a more holistic measure of school quality—and the quality of the teacher environment, specifically—may capture important aspects of the schooling enterprise which are not included in current measurement and accountability frameworks.Put another way, relationships between test scores and other school quality variables is not always strong, and specific strengths and weaknesses may be hidden even if an aggregate relationship is positive.

We find that School Culture metrics exhibited the strongest relationship to PPI, with an average correlation of moderate magnitude (ρ=0.49). Given the robust association between school climate and achievement (Thapa, Cohen, Guffey, & Higgins-D'Alessandro, 2013), such a finding was anticipated. The seven individual metrics that form School Culture—taken from the teacher survey, student survey, and district administrative data—all exhibit positive associations with state PPI, although the magnitude of these correlation coefficients varies widely. This suggests that even in the case where broader constructs exhibit a strong connection to the PPI, the metrics composing the broader construct are likely to behave differently.

INSERT TABLE 5 ABOUT HERE

On average, the metrics forming the Resources category displayed weak-to-moderate positive correlations (ρ=0.25) with PPI. Within Resources, the parental engagement and community engagement scales—both drawn from the teacher survey—exhibited strong and moderate correlations, respectively, with thestate PPI calculation. This might not be surprising, given that student achievement is strongly influenced by home and community effects that might be reflected in these engagement scales. A number of metrics within the control of the school, however, had negative correlations with PPI; these metrics were art classes per student, the support staff scale, and class size. The near-zero relationship between PPI and art classes per student is to be expected, given that PPI likely does not capture the benefits of arts education. These findings suggest that Resources, which generally reflect OTL concepts more than outcome-based indicators of school quality, are not captured very well by the PPI.

Academic Learning metrics and PPI were also weakly-to-moderately correlated (ρ=0.33). However, when one looks at metric-level correlations within the Academic Learning category, a compelling trend emerges.Three metricswere strongly correlated with PPI. The first of those, the state Student Growth Percentile (SGP), is one of four components of thePPI.Consequently, the strong correlation between the two is to be expected. Similarly, the student achievement scale—a teacher survey measure that captures perceptions about student work ethic and performance—and the Problem Solving scale—measuring teacher perceptions of student higher order thinking skills—also correlate highly with PPI. However, two metrics exhibited a negative correlation with PPI: the student engagement and valuing learning scales, both of which seek to measure student connectedness to learning. One may view these latter two metrics as reflectingan opportunity to learn, whereas the former three metrics better capture the level of student performance. Thus, this finding suggests that PPI may be capturing student performance without

doing a particularly good job of representing the extent to which teachers provide students with a chance to learn by getting students engaged in the process of their own learning. In fact, schools that perform better on student achievement might be*damaging* the intrinsic value of learning in the process.

Metrics within the Character and Wellbeing category, on average, exhibit near zero correlation (ρ=0.05) with the PPI. This is largely due to one metric—appreciation for diversity—having a strong positive correlation, and the remaining four metrics exhibiting negative correlations to PPI. Somewhat surprisingly, schools in the district exhibited very little variability on metrics within Character and Wellbeing category. In other words, while schools differ dramatically according to the PPI, they look remarkably similar when comparisons are made using SEL indicators. This may be due to the fact that the district exerts a stronger influence than the school in this domain, or it may be due to measurement error. Whatever the case, though, it seems to call for more thorough investigation.

Relationships between the Alternative Framework Metrics and School Poverty

One of the strongest criticisms of existing data systems is that they unintentionally capture out-of-school factors tied to student demography. In other words, rather than measuring schools, they are measuring families and neighborhoods. Here we answer our second research question by examining whether SQF metrics relate as strongly to school poverty as does the existing state PPI score, which is heavily reliant on raw standardized test scores.

The Massachusetts PPI calculation exhibits a very strong negative correlation (-0.80) to the percentage of economically disadvantaged (ED) students in a school, while the relationships between SQF metrics and percent ED vary in magnitude (see table 5). Overall, the average correlation for all metrics (ρ = -0.44) is roughly half as large as the correlation between PPI and

ED rates. In three major categories—Teachers and the Teaching Environment, Resources, and Character and Wellbeing—we see, on average, moderate negative correlations to ED rates. While several metrics within these categories are, in fact,tightly tied to school poverty—e.g. Professional Preparation (ρ = -0.92), Curricular Strength (ρ = -0.80), Parental Engagement (ρ = -0.76)—most exhibit more moderate correlations. In fact, five of the 20 metrics from these categories (Interest in Students, PD Scale, Art Classes per Student, Class Size Scale, Grit Scale) exhibit near-zero or slightly positive correlations, while Class Size has a positive correlation with ED rates of moderate magnitude (ρ = 0.41), representing an important investment made by the district into its poorest school. Unsurprisingly, each of these five metrics reflect OTL or SEL themes more so than absolute student academic performance. Conversely, we find School Culture metrics to be consistently and strongly tied to school poverty.

The most interesting trends to emerge from this particular analysis are seen in the correlation coefficient between ED rates and Academic Learning measures. Two metrics—the Student Achievement Scale and the Problem Solving Scale—exhibit near-deterministic relationships with the percentage of ED students in a school, with correlation coefficients of -0.99 and -0.95, respectively. The state SGP score, which, roughly speaking, measures achievement growth and not absolute achievement, still exhibits a negative correlation of moderate magnitude (ρ = -0.33). However, two metrics within Academic Learning (Student Engagement Scale, Valuing Learning Scale) have essentially no relationship to school poverty. Given that these latter two variables are closer representations of opportunity-to-learn (student perceptions of their engagement in class, teacher perceptions of whether students value learning), and further from the more absolute learning metrics that comprise this category (teacher perceptions of student ability to achieve, problem solve), this provides a compelling example of

how aholistic accountability system provides a more complete picture of quality for those schools serving vulnerable students.

Discussion

Measurement systems shape school priorities, inform policy, and affect parental behavior. They also constitute the basis for accountability structures. This study examines how the current system used to identify school quality in Massachusetts—the Progress and Performance Index—compares to a comprehensive alternative system that may prefigure accountability systems of the future. Specifically, weexaminewhat the state model fails to capture, as well as what it captures but should not. We find that the PPI calculation does roughly align with a more comprehensive framework. However, the PPI system suffers from several weaknesses previously identified by scholars, educators, and the public: it offers only summative information that cannot be used for school improvement, it fails to capture information about the opportunity to learn and social emotional learning, and it strongly reflects school demographics. By contrast, the SQF model—as an example of what accountability systems of the future might look like—is less prone to these particular shortcomings.

This study may strengthen the hands of those who have identified weaknesses in current approaches to measurement and accountability. For, while there is much agreement that current systems are inadequate, we know relatively little about the degree to which an alternate system would offer an improvement. As evidence from this study appears to indicate, a model that includes a broader range of metrics represents a significant step forward.

What is Not Measured by Current Systems

One important theme to emerge from this work is that the current accountability framework in Massachusetts—typical of such frameworks in other states—does not capture all

of the elements of school quality that stakeholders deem to be important. Although we find that PPI is positively related to each of the five measures we use to define school quality, this relationship is quite weak in some cases, and numerous important metrics are actually inversely related to PPI.Even when looking only at the metrics which comprise the Academic Learning category of the alternative SQF model, we see a range of associations which might indicate that PPI fails to capture certain components of academic achievement.There are a number of important practical and political implications that follow from this.

School Rankings

This work documents the multifaceted nature of school quality, and reveals some of the nuance lost when school quality is presented as unidimensional. In cases where ranking must be done to identify low-performing schools, it is important to note three things. First, identified schools may have different strengths and weaknesses. Second, some domains of school quality may differentiate schools quite well, while others may be relatively consistent across schools. Third, regardless of the methodology used to rank schools, one should acknowledge that rankings are highly dependent on the metrics that are chosen for inclusion, and that such choices involve a subjective component.

Policymakers, of course, do not have to rank schools. But if they are going to, such a high-stakes practice demands a more complete accounting of school quality.

Gaming

A second policy implication relates to the unintended consequences of measurement systems. Critics argue that traditional accountability systems, being heavily reliant on a single measure, may promote gaming. That is, they may encourage schools to improve their scores on a performance indicator without actually improving their overall performance. So, while it may

sometimes be the case that rising test scores indicate gains in student learning, it might also be the case that rising scores indicate a narrowing of the curriculum or an increased emphasis on test-preparation techniques—practices that would accomplish the same end by different (and problematic) means.Consequently, it appears to be in the best interest of students to create a system that is harder to game.

Since a holistic system has far more indicators, and its constituent metrics are not deterministically related to each other, it appears less likely that such gaming behaviors would develop.That said, the strongest incentive to game performance indicators may be high-stakes accountability itself, which places tremendous pressure on schools to achieve measurable results. Insofar as that is the case, policy leaders may wish to revisit not only the performance measures they include in accountability systems, but also the stakes attached to those systems.

Actionable Data

One of the goals of any accountability system shouldbe to provide useful information to stakeholders. Given that individuals may differentially value particular aspects of school quality, it seems important to report on multiple measures. A more comprehensive school quality measure is more likely to align with the interests of the public and to provide them with actionable information. It is also likely to provide schools with more information. When schools are able to see how they compare to each other along a range of metrics, as opposed to merely achievement, leaders are more apt to make judgements based on such data.

What Is (But Should Not Be) Measured by Current Systems

In addition to current accountability systems failing to measure certain aspects of schooling, they also indirectly measure family and neighborhood characteristics. Though demography is not destiny, and though schools with similar poverty levelsdo vary in their ability

to improve student outcomes, the relationship between school poverty and accountability ratings is quite strong.Two important implications flow from our finding.

Perceptions of School Quality

Current accountability systems are highly reliant on test scores. Insofar as perceptions of school quality are shaped by such systems, then, they will be strongly influenced by student poverty. Perceptions of school quality matter enormously, likely driving a subset of teachers and parentstoward higher-achieving schools and away from those identified as struggling. It seems quite possible that such a Matthew Effect would increase inequality in public education, with systems of accountability perversely hurting the very schools they were established to help.

We find variability in the relationship between SQF metrics and school poverty, and observethat, on average, the magnitude of this correlation is half as large as that between the state accountability system and school poverty.The inclusion of a more comprehensive set ofindicators, it seems—particularly if they were less tightly coupled with socioeconomic variables—mighthelp to highlight the ways in which schools serving historically marginalized groups are, in many cases, doing rather well.

Capacity Building

Under an alternative accountability system, the relative strengths and weaknesses of schools are more likely to emerge. Indeed, this analysis reveals important areas where schools with low test scores perform roughly on par with (or better than)higher scoring schools. This is not to say that we should be sanguine about low test scores. Though test scores are limited indicators of school quality, they do indicate something about basic literacy and numeracy, and by extension, about a core function of public education.For the past two decades, however, low test scores have been viewed as the sign of a failing school and have served as the basis for

sanctioning the schools most in need of assistance. Amore holistic framework, which more clearly identifies strengths and weaknesses,and which identifies inputs alongside outputs, may help to shift accountability systems awayfrom punishing schools and towardcapacity building.If it is possible to determine where and why a school is weak, and if it is clear that the school is not uniformly underperforming, it may seem less reasonable to label it failing or to slate it for closure.

It is worth noting here that, although schools with low test scores have been more effected than those with relatively high scores, a more holistic measurement system might benefit *both* groups of schools. As we find in this study, schools are not uniformly good or bad. Thus, schools with high standardized test scores may have areas of relative weakness that have been overlooked and therefore unaddressed. Seeing schools with more nuance, then, may lead to more emphasis on capacity building, whetherstudent standardized test scores are high or low.

Limitations

The data used in this study come from the initial pilot year of an initiative to create a more holistic measure of school quality,using a newframework that may be refined and expanded upon in subsequent years. There are clearly numerous concerns that must be evaluated before any new metric should be used in a high-stakes accountability system, including the benefits and drawbacks of expanding school quality measures, as well as possible unintended consequences of doing so. Additionally, no alternative accountability framework will reflect stakeholder values perfectly, even one like the SQF, which was developed in collaboration with local stakeholders. This study also examines only seven schools in a single district, all of which are subject to a single state accountability scheme. Moreover, this sample is restricted to traditional primary/middle schools, and only students in grades 4 and above were surveyed; we do not

include any early education centers, high schools, or alternative schools in our analyses. Overall, then, one should be very cautious about generalizing the findings presented here to other schools. That said, given the dearth of research on this topic, as well as largely similar incarnationsof accountability across states in the nation, the findings here are seemingly germane to a wide audience.

Conclusion

The recent authorization of ESSA will likely spur the inclusion of additional school quality metrics in measurement and accountability systems, most likely in the form of opportunities to learn and socioemotional learning. The findings in this study support the continued exploration of a more holistic measure of school quality. Current accountability systems measure too little about schools, and too much about families and neighborhoods.

Insofar as accountability systems seek to encourage efficient and effective use of resources, it seems they have much to gain from the kinds of improvements described here. But we must recall that accountability systems in education are also intended to promoteequity for our most vulnerable students, who deserve a fair and adequate education. For this task, current measurement and accountability systems appear even lessup to the task. By stigmatizing and sanctioning low-achieving schools without understanding how well such schools perform across their full mission, we exacerbateinequality of opportunity. Those harmed, as a result, are those most in need of our care.

The accountability system of the future, if it looks like what we imagine, will not be perfect. But it does represent a significant improvement. Policymakers should take seriously the challenge of moving forward and revising existing measurement and accountability systems.

And, as they do, they should remember that an even more perfect system lies even further ahead. Beyond each mountain, another mountain.

Notes:

1. Although steps are being taken to produce criteria-based measures of school quality for this district, our lens in this study is norm-referenced. Thus, both the school quality framework used here and the state PPI compare schools to each other. Given the nature of this project, we are limited to comparisons between schools within the district.
2. For more information on PPI, see http://profiles.doe.mass.edu/accountability/report/aboutdata.aspx

References

5Essentials (n.d.). Support center. Retrieved from website: http://help.5-essentials.org/customer/portal/articles/780471-illinois-5essentials-faqs

Archibald, S. (2006). Narrowing in on educational resources that do affect student achievement. *Peabody Journal of Education*, *81*(4), 23-42.

ASCD. (2014). *Multimetric accountability among ASCD's key policy priorities. Education Update*, *56*(5), 8.

Author. (2017).

Cowley, P. (2006). For a more complete report card, just add data. *Fraser Forum*, 23-31.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Berkowitz, R., Moore, H., Astor, R. A., &Benbenishty, R. (2016). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research*, 0034654316669821.

Cadima, J., Peixoto, C., & Leal, T. (2014). Observed classroom quality in first grade: associations with teacher, classroom, and school characteristics. *European Journal of Psychology of Education*, *29*(1), 139-158.

Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences*, *24*(1), 154-176.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives*,*8*(1), 1-44.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *The Teachers College Record*, *106*(6), 1047-1085.

Darling-Hammond, L. (2007). Standards and accountability movement needs to push, not punish. *Journal of Staff Development*, *28*(4), 47-50.

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., &Stosich, E. L. (2016). *Pathways to new accountability through the Every Student Succeeds Act*. Palo Alto, CA: Learning Policy Institute. Retrieved from website: http://learningpolicyinstitute.org/our-work/publications-resources/pathways-new-accountability-every-student-succeeds-act

Davis-Kean, P.E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology* 19(2), 294-304.

Downey, D.B., von Hippel, P.T., & Hughes, M. (2008). Are "failing" schools really failing? Using seasonal comparison to evaluate school effectiveness. *Sociology of Education 81*(3): 242-270.

Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (383-417). The Netherlands: North-Holland.

Grissom, J. A., Loeb, S., & Doss, C. (2015). The multiple dimensions of teacher quality: does value-added capture teachers' nonachievement contributions to their schools? In *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures*, 37.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, *19*(2), 141-164.

Hanushek, E. A. (2003). The Failure of input-based schooling policies. *The Economic Journal*, *113*(485), F64-F98.

Houtenville, A. J., & Conway, K. S. (2008). Parental effort, school resources, and student achievement. *Journal of Human Resources*, *43*(2), 437-453.

Jackson, J. J., Connolly, J. J., Garrison, S. M., Leveille, M. M., & Connolly, S. L. (2015). Your friends know how long you will live: A 75-year study of peer-rated personality traits. *Psychological Science*, *26*(3), 335-340.

Jackson, C. K. (2016). *What do test scores miss? The importance of teacher effects on non-test score outcomes. Working Paper No. 22226.* Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., &Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper*. MET Project. Bill & Melinda Gates Foundation.

Kutsyuruba, B., Klinger, D. A., & Hussain, A. (2015). Relationships among school climate, school safety, and student achievement and well-being: a review of the literature. *Review of Education*, *3*(2), 103-135.

Ladd, H., & Loeb, S. (2013). The challenges of measuring school quality. In D. Allen and R. Reich (Eds.),*Education, Justice, and Democracy* (19-42). Chicago, IL: The University of Chicago Press.

Lavy, V. (2012). *Expanding school resources and increasing time on task: Effects of a policy experiment in Israel on student academic achievement and behavior*. *Working Paper No. 18369.* Cambridge, MA: National Bureau of Economic Research.

Lubienski, S. T., Lubienski, C., & Crane, C. C. (2008). Achievement differences and school type: The role of school climate, teacher certification, and instruction. *American Journal of Education*, *115*(1), 97–138.

Markow, D. &Pieters, A. (2012). The MetLife survey of the American teacher: Teachers, parents and the economy. New York: MetLife, Inc.

Massachusetts Department of Elementary and Secondary Education (MDESE). (2016). School and district profiles. Retrieved from http://profiles.doe.mass.edu/

Mintrop, H. &Sunderman, G.L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement—and why we may retain it anyway. *Educational Researcher 38*(5): 353-364.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., &Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2), 13-22.

National Educaiton Assocaition (NEA). (2011). Beyond two test scores: Multiple measures of student learning and school accountability. An NEA policy brief. Retrieved from website: http://www.nea.org/assets/docs/PB38beyondtwotestscores2011.pdf

National Council of Teachers of English (NCTE). (2012). Opportunity-to-learn standards, statement of principles. Retrieved from http://www.ncte.org/positions/statements/opptolearnstandards

Nunnaly, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Petek, N., & Pope, N. G. (2016). The multidimensional impact of teachers on students. Working paper. Retrieved from http://home.uchicago.edu/~npope/Nolan_Pope_JMP.pdf

Reardon, S.F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G.J Duncan & R.J. Murnane (Eds.), *Whither Opportunity?* New York: Russell Sage Foundation.

Rothstein, R. & Jacobsen, R. J. (2006). The goals of education. *Phi Delta Kappan*, *88*(4), 264-272.

Rumberger, R. W.,& Palardy, G. J. (2005). Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal*, *42*(1), 3–42. doi:10.3102/00028312042001003

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., &Barterian, J. A. (2013). Heighted test anxiety among young children: Elementary school students' anxious responses to high-takes testing. *Psychology in the Schools*, *50*(5), 489-499. doi:10.1002/pits.21689

Spalding, A. (2014). *The 2014 Michigan public high school context and performance report card*. Midland, MI: Mackinac Center for Public Policy.

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, *83*(3), 357-385.

United States Education Department (USED). (2013). *State and local report cards. Title I, Part A of the Elementary and Secondary Education Act of 1965, as amended. Non-regulatory guidance*. Retrieved from website: http://www2.ed.gov/programs/titleiparta/state_local_report_card_guidance_2-08-2013.pdf

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., &Maughan, R. (2000). Validation of student, principal, and self-ratings in 360 Feedback® for teacher evaluation, *Journal of Personnel Evaluation in Education*, *14*(2), 179-192.

Appendix A: School Quality Framework (SQF)

ESSENTIAL INPUTS

1. TEACHERS AND THE TEACHING ENVIRONMENT

- 1A. Knowledge and Skills of Teachers
  - 1Aic. Professional Preparation Scale
  - 1Aiia. Pedagogical Effectiveness Scale
  - 1Aiiia Interest in Students Scale
- 1B. Teaching Environment
  - 1Bia. Teacher Turnover
  - 1Biib. Professional Development Scale
  - 1Biiia. Teacher Principal Trust Scale
  - 1Biiib. Principal Instructional Leadership Scale

2. SCHOOL CULTURE

- 2A. Safety
  - 2Aib. Student Safety Scale
  - 2Aiib. Peer Victimization Scale
  - 2Aiic. Peer Support Scale
- 2B. Relationships
  - 2Bia. Sense of Belonging Scale
  - 2Biia. Student Teacher Relationship Scale
- 2C. Academic Orientation
  - 2Cia. Chronic Absences
  - 2Ciia. Academic Press Scale

3. RESOURCES

3A. Facilities and Personnel

3Aiia. Art Classes per Student

3Aiib. Counselors per Students

3Aiid. Support Staff Scale

3B. Curricular Resources

3Bif. Curricular Strength Scale

3Biia. Class Size

3Biib. Class Size Scale

3C. Community Support

3Cia. Parental Engagement Scale

3Ciia. Community Engagement Scale

KEY OUTCOMES

4. INDICATORS OF ACADEMIC LEARNING

4A. Performance

4Aia. State SGP Score

4Aiia. Student Achievement Scale

4B. Student Commitment to Learning

4Bia. Student Engagement Scale

4Biia. Valuing Learning Scale

4C. Critical Thinking

4Cia. Problem Solving Scale

4D. College and Career Readiness

5. CHARACTER AND WELLBEING OUTCOMES

5A. Civic Engagement

5Aiia. Appreciation for Diversity Scale

5B. Work Ethic

5Bia. Grit Scale

5C. Artistic and Creative Traits

5Cia. Arts Exposure

5D. Health

5Dia. Positive Affect Scale

5Diia. Physical Activity