Vital Signs for the American Education System

Jennifer L. Jennings
Princeton University

Nicholas Mark
New York University

In this paper, we propose a set of indicators to track between-group inequalities in K-12 educational outcomes and experiences. In Section I, we begin by providing a conceptual framework for indicator system design and use. We classify indicators into four types (*outcome, direct, enabling, and diagnostic indicators*) and discuss their strengths and weaknesses for making particular types of inferences. By identifying multiple user types (*policymakers, researchers, practitioners, and the public*) and considering the manifold potential uses of indicators (*evaluative, descriptive, and predictive*), we argue that an important first task for indicator system designers is to clarify the system's end users, purposes, and theory of change.

In Section II, we describe the major dimensions of academic achievement, progress, and engagement. Doing so requires a dual frame. We review the construct definitions used in existing research, while also noting the ways that these constructs fall short in capturing the types of achievement and engagement required by a labor market that values non-automatable skills. Put simply, the academic and engagement skills that will be important in the future are not necessarily those that have been important in the past. It is useful to note, however, that concerns about a mismatch between available measures and researchers' desired conceptualization dates back to indicator discussions in the late 1980s (Oakes 1989). This is, in some respects, an

intractable problem in an era of rapid social and economic change, but one that is worth acknowledging at the outset.

In Section III, we focus our attention on inputs and processes that are plausibly associated with outcome disparities in K-12 education by income and race/ethnicity: principal leadership, teacher characteristics, student composition, access to curricular pathways, and opportunities for engagement (attendance and extracurricular activities). We emphasize that for any input or process to contribute to group disparities, groups must have either *differential exposure* to that input, *differential sensitivity* to that input (i.e. the effects of input *x* on outcome *y* varies across groups), or both. Across each of the input categories considered, we find that the evidence on the average effects of these characteristics is often more comprehensive than the evidence on their heterogeneous effects across groups. Finally, a major challenge for indicator development is that studies that allow for causal inference about the relationship between inputs and outcomes tend to evaluate impacts on students' state test scores; few directly evaluate the impacts of these inputs on forward-looking measures of achievement and engagement.

In Section III, we propose a set of indicators, while acknowledging the gaps in our knowledge about the effects of these inputs. We reflect on the proliferation of indicator systems in the last half century both domestically and abroad, and focus on two design considerations (*information overload* and *local salience*). In doing so, we contend that a small set of diagnostic and enabling indicators, akin to vital signs, that can be produced at the level of states, congressional districts, state assembly and senate districts, school districts, and individual schools are preferable to a set of indicators that are likely to be prohibitively expensive to collect at any level beyond a nationally representative sample.

As with the use of vital signs in medicine, the indicators we propose are intended to provide a quick and manageable characterization of the health of key parts of the educational system, rather than to comprehensively describe the health of *all* parts of the education system, and call attention to trends in educational conditions that deserve further investigation. Our recommendations, however, diverge from historical practice by calling for cumulative exposure in addition to cross-sectional measures. While we are currently aware of no other systems that do so, we argue that this approach is more consistent with describing differential group exposure to inputs and processes that collectively affect their short, medium, and long-term educational success.

I. Indicator Systems: Design Decisions

History and Context

As chroniclers of the social indicator movement have pointed out, indicator systems have many potential purposes and audiences, which we review below. Our purpose in briefly describing this history and concomitant writing about indicators is to emphasize that settling on a purpose and end user is a necessary – though often unstated - component of indicator system creation. In addition, we draw readers' attention to the longevity of the indicators debate in education, and suggest that it is worthwhile to consider whether prior indicator systems have been effective at achieving their stated goals, and, if not, to understand why.

From its outset in the 1960s, observers of the social indicators movement debated whether their purpose should be understood as *enlightenment* or *national goals accounting* (Land 1975). "To contribute to the enlightenment of a changing society," as Duncan (1974) wrote, was the optimal approach for social indicators, providing an understanding of general social patterns, rather than to serve as "philosopher kings…creating the best managed society that knowledge can design." In contrast, others operating from a *goals accounting* perspective had no qualms about assuming the normative mantel with which Duncan was concerned. For example, Olson (1969) saw indicators as a "statistic of direct normative interest which facilitates concise, comprehensive and balanced judgments about the condition of major aspects of a society. It is, in all cases, a direct measure of welfare and is subject to the interpretation that if it changes in the 'right' direction, while other things remain equal, things have gotten better, or people are better off" (Olson, cited in Land). But, as many observers have pointed out, such a definition requires some level of consensus about the measures of welfare worth improving as well as agreement about what it means to be "better off."

Later work on social indicators focused on the ways that indicator choice fundamentally frames policy questions and the terms of debate rather than providing the answers to these questions themselves (Innes 1989), and changes problem definition (Cobb and Rixford 1998). By drawing attention to one set of indicators versus others and highlighting their interrelationships, these authors argued, some problems are amplified while others are deemphasized. Measures themselves – the practice of tracking statistics in particular ways – both reflect and transform the meanings of the constructs being measured (Scott 1998).

Educational indicator systems in the United States have been debated at least since the 1980s. As Smith (1988: 487) quipped, by the end of the 1980s, nearly every educational group or agency was in the indicator creation business. As in the broader debate about social indicators, from the outset education scholars debated to what purposes these systems should be put. Murnane (1987) argued for an exclusive focus on outputs. Smith (1988) argued for predictive function of indicators, seeing the value to policymakers as allowing them "to anticipate changes in educational outcomes, rather than merely to describe them" (p. 488). In contrast, Guthrie (1993) argued for a "Dow Jones Index"-like composite indicator for education that could mobilize regular attention around educational issues. Darling-Hammond (1992: 236) drew attention to the issue of educators' reactivity to measurement, noting that the goal should be to "illuminate educational issues without distorting the educational process." Elsewhere, Oakes (1993) argued for the importance of integrating measures of context into indicator systems, noting, "We can't link many school characteristics directly to student learning and persistence in school. Nevertheless, they do either enable or constrain teaching and learning in classrooms" (p. 183). Still others, like Bryk and Hermanson (1993) argued that the test of the success of indicator systems should be evaluated against their ability to "enrich public discourse" rather than their instrumental purposes: specifically, whether the new information "actually deepens understandings and stimulates a more extended discourse around the aims and means of education" (p. 465).

In short, by the early 1990s, there was no shortage of opinions about the purposes and challenges of indicator system creation in a decentralized education system. School outputs ultimately overtook input and process indicators, with states and ultimately the federal government adopting accountability systems based on student test scores. We provide this context to emphasize that 30 more years of educational research has not settled these debates. While we have learned much in three decades, Oakes' (1993) statement about the inability to unequivocally link school characteristics to student learning and persistence still provides an accurate characterization of the state of our knowledge about most inputs' effects.

Indicator Systems Uses

We divide indicator use into *predictive*, *descriptive*, and *evaluative* uses. *Predictive indicators* provide an "early warning" system, identifying issues, organizations, or individuals

that need additional attention (Allensworth, Nagaoka, and Johnson 2018). In contrast to other types of indicators, these are intended to be produced and reviewed multiple times per school year in order to guide school improvement and student support efforts. The most commonly used predictive indicator in K-8 education is arguably benchmark assessment, which is intended to predict performance on summative assessments and allow educators to reteach or allocate extra attention to students struggling in a particular area. (Goertz et al. 2009; Halverson 2010). In addition to benchmark assessment, high schools often use "on time graduation indicators," such as that used in the Chicago Public Schools (Allensworth et al. 2018). Predictive indicators must be provided "just in time" to address the issues that they uncover.

Descriptive indicators are those that identify change over time along outcomes of interest. They are useful for researchers and policymakers to track outcomes over time and across groups or organizations. More generally, they provide the public with the means to make sense of the current state of the education system. Descriptive indicators do not try to attribute responsibility for an outcome to a particular actor or organization, but offer an overall assessment of the health and functioning of some part of the system. This category, for example, includes recent National Assessment Governing Board efforts to use student and educator NAEP surveys to produce a set of 10-15 key educational indicators (Ginsburg and Smith 2014).

Evaluative indicators are used to make inferences about whether an organization is effectively fulfilling its responsibility to its clients. At least since the 1990s, states have relied on state tests as evaluative indicators for schools, and, in more recent years, for teachers. Evaluative indicators are typically those described as outcome indicators in the typology offered below, although there are no theoretical reasons why they could not monitor inputs and processes as well.

Indicator Types

Having characterized their main uses in the prior section, we divide indicators into four different types: direct, outcome, enabling, and diagnostic. Each of these four types differs with respect to the breadth of the construct it measures, the effect that changing the indicator has on an outcome, and the effect, sometimes unintended, that changing the indicator would have on other outcomes.

Direct indicators measure inputs that have a direct relationship with an outcome. These are analogous to drugs that have a consistently replicated average treatment effect via known biological targets or pathways. They measure self-contained, specific inputs, such as milligrams of acetaminophen ingested, that have an effect on a specific outcome (e.g. fever). The effects of a change in a direct indicator are predictable and relatively consistent across patients. While there is a distribution of treatment effects of any drug, such that the treatment effects for an individual are unknown, at the population level, we expect a given change in a direct indicator to be reliably associated with a change in outcome *y*.

Two important dimensions of direct indicators, as we defined them here, are that social contexts do not change the effect of the input, and the effects of the input on any student are not dependent on peers' reaction to the input. It is difficult to imagine any input in education that resembles a direct indicator, since fidelity of implementation varies across schools and districts, and the effect of an input may change once it is implemented at scale. For example, whole-school reform models such as Success for All have had markedly different treatment effects in experimental and at-scale settings (Borman et al. 2003).

While we could not identify any truly direct indicators that exist in education, many inputs are discussed as if their impacts operated in this way. Value-added measures of teacher quality, for instance, attempt to isolate the effect of teacher quality on student outcomes. While these measures capture some component of teacher quality that has a direct effect on student outcomes, that effect is likely to vary between contexts and across types of students. An effective high school teacher in one school may not continue to be so in a different context. Similarly, a bilingual teacher may be particularly effective with students learning a second language, but offer no particular benefit for English-dominant students. As a result of the narrowness of the construct that direct indicators measure, they are necessarily limited to the same outcomes in similar settings.

Outcome indicators are measurements of desired outcomes themselves. Depending on the user, they can simply be used to track change over time (*descriptive use*), or may incentivize action towards a particular outcome without prescribing the process by which that outcome is achieved (*evaluative use)*. The outcome may be broad, such as BA attainment or life expectancy, or narrow, such as filing a FAFSA form. Broad outcomes indicators are less effective at motivating behavior change, and because any form of measurement will shape behavior, even in

the absence of formal incentives around the indicator, this may lead to behaviors that invalidate the inferences that users hope to make using a given indicator (Campbell 1979; Koretz 2008). By remaining aware of this principle while designing indicator systems that incorporate outcomes, architects can forestall some of the issues created by individuals' and organizations' reactivity to measurement.

Enabling indicators measure inputs or processes that do not have a direct effect on an outcome, but create conditions that facilitate or constrain the effectiveness of other inputs. Enabling indicators are measurements of context, such as the availability of a math course sequence, that is related to other possible outcomes, such as the completion of a course sequence. Enabling indicators often have multiple purposes and effects. The presence of an art studio in a school, for example, may affect the number of students enrolled in art courses, but it may also affect student aspirations.

In contrast to direct indicators, which have predictable one-to-one relationships with outcomes, enabling inputs affect outcomes through multiple pathways. For example, Bryk et al. (2010), studying the relationships between organizational supports and improvement in Chicago schools, focused on five enabling indicators: the coherence of the instructional guidance system, the professional capacity of the school, the parent-community-school ties, the learning climate, and the leadership's focus on driving change. Bryk et al. emphasized that the inputs interact with each other; each is necessary for the others to be impactful. Indeed, Bryk et al. stress that the five supports form "an organized system of elements in dynamic interaction with one another," rather than representing five independent inputs.

Diagnostic indicators are those, like vital signs, that highlight the existence of a potential problem. In the medical field, vital signs (i.e. blood pressure, pulse, respiration rate, and temperature) are markers of underlying health issues, but provide little information about the specific cause of the problem. Monitoring vital signs allows physicians to track the health of a patient inexpensively and regularly, and may provide a high-level measure of the effectiveness of a treatment, but vital signs, on their own, do not indicate the solution to a problem. For example, an abnormally high temperature is an indication that a patient is unwell, but on its own a high temperature does not indicate whether a patient has the flu or is severely dehydrated. In other words, there is no one-to-one relationship between a vital sign and a specific diagnosis.

Diagnostic indicators do not stipulate a mechanistic relationship between the indicator itself and the outcome, but instead highlight focus areas to be examined in further depth. Segregation, for example, has direct effects on students' contact with diverse sets of peers, enables between-group inequality in school quality by separating groups of students, and is an indication that broader inequalities exist and likely manifest themselves in other K-12 educational outcomes like test scores, graduation rates, and college matriculation. Employing a diagnostic indicator thus implies a broad view of the purpose of indicators and a more localized commitment to understanding mechanisms and effects.

II. Dimensions of Academic Achievement, Progress, and Engagement

Broadly speaking, the goals of public education are to provide students with the skills to be contributing adults, participate in democratic and cooperative processes, and flourish in their own personal lives (Brighouse 2006). In what follows, we focus on a subcomponent of those goals: academic achievement, progress, and engagement. What these constructs mean is clearly temporally and socially bound. Basic academic skills may have sufficed in a different labor market, but increasing automation of routine tasks means that the skills that students need to be economically successful as adults are those that cannot easily be replicated by machines (Autor et al. 2003; Murnane and Levy 1996). Similarly, engagement conceptualized as compliance may have sufficed in an earlier labor market, but in an economy that increasingly gives more autonomy to skilled workers, a different type of cognitive engagement may be required.

Defining Academic Achievement: Knowledge and Skills

Standardized test scores are the most common measure researchers have used to make inferences about students' academic achievement (what students know and can do), their academic growth, and between-group disparities. [Judy comment: this is how they're measured, not a definition of the skills.] Tests are necessarily based on a sampling principle, whereby testing students on a sample of questions from a broader domain of knowledge allows us to make valid inferences about their overall level of skill and knowledge. This requires that a test designer enumerate the domain that is the target of inference, such as the Common Core State Standards

or the National Assessment of Educational Progress Framework for Mathematics.

Many scholars have noted that existing standardized test measures of achievement often fail to capture the academic skills necessary for economic success in the future. [Judy comment: Yes, but what about indicators of academic achievement? Authors seem to have slipped past this.]

For example, the 2012 National Academy report *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century,* draws attention to the importance of "deeper learning", defined as "the process through which an individual becomes capable of taking what was learned in one situation and applying it to new situations (i.e., transfer)." Similarly, Murnane and Levy (1996) argued that "the skills taught by schools today, even if test scores measure them as improving, are not sufficient for high school graduates in an increasingly-changing, challenging workplace." They introduce new "basic skills", including the ability to do reading and math at the 9th grade level or higher, the ability to solve semi-structured problems where hypotheses must be tested, the ability to work in groups with persons of different backgrounds, the ability to communicate effectively, both orally and in writing, and the ability to use computers to carry out tasks.

While knowledge and skills as measured by tests are positively associated with adult earnings, they nonetheless explain a small fraction of the variation in adult income and job evaluation. For example, NRC (1989) synthesized data from 700 studies relating measures of cognitive ability and job performance, and found that test scores explained 6 percent of the variance in performance. More recent studies, such as Borghans et al. (2016), also find that achievement tests alone predict approximately 6 percent of the variation in adult wages.

Non-test Measures of Achievement and Progress: Grades, grade promotion, and credit accumulation.

While grades and achievement test scores in the same subject are positively correlated, typically between .5 and .6 (Duckworth, Quinn, and Tsukayama 2012; Pattison, Grodsky, and Muller 2013), grades also measure constructs beyond subject-specific academic knowledge and skills, such as self-control and task persistence. Grades are also strongly associated with high school graduation (Allensworth 2013; Bowers 2010), as well as students' odds of graduating from college (Bowen, Chingos, and McPherson 2018); in general, grades appear to predict

college success as well as, or better than, test scores. Grades alone also predict adult wages almost as well as achievement tests (Borghans et al. 2016).

Engagement: Behavioral, Emotional, and Cognitive

Recent work on engagement consistently describes it as a multidimensional construct reflecting behavioral, emotional, and cognitive engagement (Fredricks et al. 2004). The discussion below makes clear multiple measurement and conceptualization issues in this field, which we summarize here. First, engagement is sometimes described as a trait-like feature of individuals, while at other times is conceived as a state that is strongly affected by context (Sinclair et al. 2003). It is rare for researchers to have access to measures of a given student's engagement in different contexts at the same point in time; rather, it is more typical to model longitudinal measures in which both time has passed, and contexts have changed. This makes it difficult to decompose engagement into components that are stable within persons versus those that are effects of disparate contexts.

Second, a major challenge in the engagement literature is that researchers do not use consistent definitions across papers, nor do they consistently use the term "engagement." Researchers vary in the extent to which they bucket engagement as an overarching construct, or attempt to break it down into the constituent parts described below. Moreover, across disciplines, researchers use "big bucket" terms such as "non-cognitive skills" or "social skills" as a stand-in term for skills that fit under the behavioral, emotional, and cognitive components of engagement. A related issue is that some disciplines adopt a negative conceptualization of engagement, assuming it is the residual not explained by test scores (Heckman and Rubenstein 2001; Heckman, Stixrud, and Urzua 2006; Mueser 1979). This makes it challenging to build a cross-disciplinary cumulative knowledge base. Most of these studies use convenience constructs of behavioral and cognitive engagement, making it difficult to compare across studies and build a cumulative body of knowledge.

Third, scholars differ on what constitutes indicators of engagement (features that define the construct of student engagement), facilitators of engagement (contextual factors that affect engagement), and outcomes of engagement (outcomes predicted by engagement) (Lam et al. 2014). Depending on the research question, one can make an argument for any of these conceptualizations.

With these caveats, behavioral engagement includes "positive conduct" like attending school and completing assigned work, avoiding disruptive behavior, and participating in extracurricular activities (Wang and Eccles 2013). Behavioral engagement is typically captured using administratively captured measures such as attendance, tardiness, and suspensions; self-reports of engagement in extracurricular activities; or teacher ratings of students' behavior in class.

Emotional engagement denotes positive affective school relationships and sense of belonging. In many cases, these are self-reported constructs; in others, teachers and parents rate students on their perceptions of students' emotional engagement. As described above, researchers disagree about whether a given measure should be assumed to be a feature of a context (like student/teacher relationships), an outcome of engagement, or an indicator of engagement.

Cognitive engagement includes motivation to learn and self-regulation in the learning process. In psychology, cognitive engagement appears to fly under multiple different terms, including cognitive self-regulation, executive function, and task persistence (Blair and Diamond 2008; Blair and Razza 2007; Duckworth and Seligman 2005; Duncan et al. 2007; Raver 2004).

III. Inputs and Processes Associated with Inequality in Outcomes

In this section, we identify four broad input categories that are associated with between-group inequality: principal quality, teacher characteristics (experience and race/ethnicity/gender match), school composition, and gateway courses and curricular rigor. For each category we review the literature illustrating the connection between the input and student achievement, progress, and engagement outcomes. At present, most of the existing literature linking school inputs and processes and student outcomes relies on standardized tests of reading and math, with the exception of some measures of teacher quality. As a result, the literature that follows is dominated by these outcomes, but we integrate those using a broader set of measures where available. In addition, research has typically focused on identifying direct effects of each of these factors on student outcomes, and our review reflects this focus. However, the value of these inputs as indicators is not limited to direct effects, as we noted when we discussed enabling indicators; we return to this issue when we recommend indicators in Section IV.

Before turning to the literature, we note two challenges facing the committee. First, while the committee's statement of purpose includes disparities by socioeconomic status, race/ethnicity, gender, disability, urbanicity, and English language proficiency and, to be sure, there are inputs that influence most of these categories, some group disparities are more or less affected by particular inputs. Contrasting the cases of race and gender disparities is useful in clarifying this point. Because boys and girls are generally uniformly distributed across districts, schools, and classrooms, they are exposed to similar levels of most inputs, while spatial segregation by race means that the inputs allocated to white and black students are quite different. While *differential exposure* to a given input is more likely associated with racial disparities, *differential sensitivity* to that input is more relevant for gender disparities. For example, multiple studies suggest that boys are more sensitive to context than are girls (Legewie and DiPrete 2012; DiPrete and 2006), meaning that even in the absence of disparities in exposure to high-poverty schools, boys' achievement is more negatively affected than is girls'. Second, the existing research base is not equally strong across groups and outcomes, and it is often the case that the studies that allow us to make causal inferences about the relationship between an input and achievement or engagement outcomes use narrow outcome measures (i.e. state test scores). In addition, there are fewer studies that allow for causal inference about the impacts of inputs on measures of engagement other than achievement.

Principal Quality

Principal quality is a prime suspect in producing disparities across groups because subgroups face meaningful differences in exposure to principals with different levels of experience and turnover. Nationally-representative data from the Principal Follow-Up Study of 2012-13, for example, found that in schools in which 75% or more students qualified for free and reduced-price lunch, 72.6% of principals remained in the school in the subsequent school year, whereas 80% did in schools with fewer than 35% FRPL students (Goldring and Taie 2014). Additionally, principals at schools with more than 75% of FRPL students had an average of 3.6 years of experience at their school and 6.1 total years of experience, compared to 4.3 and 7.1 years for principals of schools with fewer than 35% FRPL students (Taie and Goldring 2017). These differences, at first, appear small. It is important to remember that they only represent

exposure in any given year rather than cumulative exposures across students' educational careers, which are likely larger.

Analyses of individual school districts provide some insight into how these patterns arise. Beteille, Kalogrides, and Loeb (2011) use longitudinal data from a large urban school district in the United States and find that principals depart more often from schools with high concentrations of students in poverty, and that they tend to move to schools with lower concentrations. New principals in schools with high concentrations of poverty are also more likely to have low levels of experience. They call this phenomenon a "stepping stone" effect, whereby principals use high-poverty schools as a first assignment before moving on to less challenging ones. Cullen and Mazzeo (2007), Branch et al. (2008), Clark et al. (2009) and Miller (2009) all found that principal turnover is higher in low-performing schools, and that principals are more likely to move to higher achieving schools when they exit a low performing one. Gates et al. (2006) found that principals were more likely to leave schools with higher proportions of minority students. Horng, Kalogrides and Loeb (2009) found that principals in high-poverty, non-white schools had less experience, lower levels of degree attainment, and attended less selective colleges. This distribution is partially driven by initial match, but is exacerbated by later movements. Principals report that these schools are hard to work in, and that they have preferences for easier work assignments. As a whole, this body of evidence demonstrates that students in schools with lower average test scores and high proportions of non-white and poor students are more likely to be exposed to more principals and higher levels of turnover.

This finding is important because the available empirical evidence points to principals as an important lever for school success, and that principal experience is closely related to student test score growth. Clark, Martorell and Rockoff (2009) use fixed effects models and data on principals and students in New York City from 1999-2007. They find evidence that principal experience is related to school performance as measured through student test scores, particularly for math test scores, and student absences. They find that the relationship between experience and student outcomes is particularly pronounced over the first few years of principal experience. Dhuey and Smith (2013) measure individual principal effects on gains in student math and reading achievement in Canada. By tracking teachers across schools and employing principal fixed effects, they estimate that a one standard deviation improvement in principal quality can boost student performance by approximately 0.3 standard deviations in both math and reading.

Coelli and Green (2012) took advantage of rotations of principals between schools in British Columbia, Canada to identify the effect of individual principals on the progress of students in the schools they led. They found that principals had heterogeneous effects on student achievement, but that the principal's influence was often much noisier in the first year than when measured over multiple years.

Two seminal articles stand out in laying the conceptual groundwork for how and why differences between principals affect student outcomes. Warren Little (1982) emphasized their central role in managing the school as a workplace and identified two ways they influenced the work environment: by fostering "collegiality" and "experimentation." Collegiality allows teachers to learn from their fellow teachers and lean on them for guidance or assistance, increasing the collective efficacy of the school as a whole. Experimentation is important because it is a way to create new solutions to problems, central to implementing any reform strategy, and continuous improvement requires that teachers feel comfortable trying new techniques and challenged to constantly improve. Hellinger and Heck (1998), extending this work, conceived of principal influence in four dimensions: the school's purposes and goals, the structure and social network of the teaching staff, the selection of the teachers themselves, and the organizational culture of the school. These dimensions constitute the enabling power of principal quality: principals influence the conditions under which teachers and students work.

Of the four general directions that Hellinger and Heck (1998) identified, the available evidence points to the third, the selection of teachers, as having the most important direct effects. Using an OLS model of student performance on cognitive tests with controls for a set of potential principal contributions, Brewer (1993) found that the largest amount of the variation could be explained by the percentage of the faculty appointed by the principal, implying that principals have an impact on student achievement through the selection of teachers. Using multiple models to isolate the effects of principals in Texas, Branch, Hanushek, and Rivkin (2012) find wide variation in principal quality in terms of their effect on test scores, with wider variation among high-poverty schools than low-poverty schools. They find a negative relationship between teacher turnover
and grade-level value added, which increases monotonically as principal quality rises. They conclude from this pattern that the most important role of principals is "management of the teacher force." Earlier findings from Beteille, Kalogrides, and Loeb (2011) also support this

conclusion. Their results, using data from Miami-Dade County Public Schools from 2004 through 2009 and student fixed effects models along with a rich set of controls, indicated that the higher levels of principal mobility experienced in high-poverty schools causes higher teacher turnover, which leads to lower student academic achievement gains. Most recently, Fryer (2017) conducted a school-level randomized field experiment in which principals received 300 hours of training on lesson planning, data-driven instruction, and teacher observation and coaching. They found positive effects on test scores in the first year but no effect in the second year. They find that their intervention compares quite favorably with others in terms of cost-effectiveness, such as class size reductions.

On the other hand, researchers have found that principals play other important roles as well. Grissom and Loeb (2011) conducted a survey of principals, assistant principals, parents, and teachers in Miami-Dade County and found that "organization management" was the most important predictor of principal effectiveness. Those skills included hiring personnel, but also included managing budgets, dealing with staff concerns, and developing a safe school environment. Also using survey data, Supovitz, Sirinides, and May (2010) found that principal activity was significantly associated with the extent to which teachers report larger advice networks and constructive interaction with their peers.

Recent work by Burkhauser (2017) indicates that changes in the teacher force may not be completely attributable to conscious decisions by the principal, but might rather be a result of the teachers' reactions to the principal's effect on the school. She points to the ways in which principals shape working conditions for teachers through time use and organizational practices, which is associated with the school climate and subsequent teacher turnover. Boyd et al. (2011), using survey and administrative data from New York City, find that teachers' perceptions of the school administration have a larger effect on their decision to continue teaching in that school than any other factor. Grissom (2011) also used survey data and found that ratings of principal effectiveness are associated with greater teacher satisfaction and their probability of remaining at the school.

Together, these studies suggest that principals influence on the quality of education their students receive, which teachers are responsible for educating their students, and the conditions under which those teachers work. However, like the studies of teachers we describe below, while

there is consensus that principals vary in quality, we are aware of no observable characteristics of principals associated with better outcomes for students.

Teacher Characteristics

Teachers have long been identified as key to students' academic and social progress, and also as a potential source of inequality in outcomes (Isenberg et al. 2013). In this section, we review the literature on teacher effects on both academic achievement and the body of skills not measured by test scores that coincide with cognitive, behavioral, and emotional engagement constructs. We divide this literature into studies that look at variation in students' exposure to, and the impact of, *teacher observable credentials*, and contrast these findings with those that directly estimate *teachers' value-added* on academic and other outcomes.

Research on teacher credentials has largely determined that little of the variation in teacher effectiveness can be explained by traditional measures of teacher quality, such as certification and degree attainment (Hanushek and Rivkin 2010; Aaronson, Barrow, and Sander, 2007; Kane, Rockoff, and Staiger, 2008; Rockoff 2004). It is the case, however, that students vary substantially across racial and socioeconomic subgroups in their exposure to teachers who are certified, experienced, and who majored in the subject they are teaching (Corcoran and Evans 2008). Goldhaber et al. (2015) found a similar pattern in Washington State, where "every measure of teacher quality – experience, licensure exam score, and value-added estimates of effectiveness – is inequitably distributed across every indicator of student disadvantage," including racial, socioeconomic, and ability categories (p. 12).

In the last two decades, scholars have used value-added methods (VAM), which essentially estimate the mean difference in students' observed versus predicted performance based on prior test scores and other demographic characteristics, to isolate individual teacher effects (Kane and Staiger 2008; Nye, Konstantopoulos, and Hedges 2004; Chetty, Friedman, and Rockoff 2014; Hanushek and Rivkin 2010; Rivkin, Hanushek, and Kain, 2005; Koedel, Mihaly, and Rockoff 2015). In general, these studies find that a standard deviation in the teacher effectiveness distribution (as measured using value-added methods) is associated with a .1 standard deviation increase in test scores. Other research has now examined the effects of higher value-added teachers on a range of life outcomes. For example, Chetty, Friedman, and Rockoff (2011) found that students assigned to teachers near the top of the value-added distribution are

better off than students assigned to teachers near the bottom on a variety of outcomes, including likelihood of attending college, attending higher-ranked colleges, earning higher salaries, living in higher SES neighborhoods, saving more for retirement, and becoming teenage parents. While many critiques of value-added methods and their potential uses in selecting and retaining teachers exist (Harris, 2011), they nonetheless have clarified that teachers vary considerably in their contribution to students' outcomes.

Other studies have attempted to isolate teacher effects in other ways. The Gates Foundation's Measures of Effective Teaching (MET) project, which concluded in 2013, set out to do so by randomly assigning students to teachers and using multiple measures of teacher effectiveness. The study was a multi-year venture encompassing thousands of teachers in six school districts across the country. They found that the direction and magnitude of teacher effects on student achievement growth were predictable based on prior observations, confirming the idea that teachers do vary in the extent to which they influence student performance and that that variation is both measurable and predictable (Kane et al. 2013). The researchers also attempted to use rich observational and value-added data to estimate a "composite estimator of effective teaching" (Mihaly et al. 2013), but found that the data did not support the use of a single estimator of effective teaching (Rothstein and Mathis 2013). Instead, they found that although there was a relatively stable component of effectiveness, other components were far from perfectly correlated with this stable measure (Mihaly et al.2013).

Test scores are only one component of academic achievement, and do not provide an independent measure of engagement. A second set of studies have focused on teacher influence on behaviors that generally fall under the rubric of engagement, but, as noted earlier, are referred to using different labels across disciplines. This research has documented the effects of teachers on behavior (Ladd and Sorenson 2015; Jackson 2014, Blazar and Kraft 2017), absences (Ladd and Sorenson 2015; Jackson 2014), complex cognitive skills (Kraft and Grace 2016), social and behavioral skills (Jennings and DiPrete 2010; Kraft *forthcoming*), graduation (Koedel 2008; Jackson 2014), motivation and self-efficacy (Ruzek et al. 2015; Blazar and Kraft 2017), grade progression, and grades (Jackson 2014). Existing research has found only moderate to weak relationships between teacher effects on achievement and measures that fall more consistently under the rubric of engagement (Jackson 2014; Jennings and DiPrete 2010; Kraft and Grace, 2016; Kraft *forthcoming;* Blazar and Kraft 2017); knowing that a teacher is high value-added on

test score outcomes provides little information about their performance on outcomes that serve as proxies for engagement.

Though most teacher observable characteristics are weakly correlated with teachers' value-added to academic and social skills, teacher experience, particularly the distribution of novice teachers, is the one credential-related characteristic that is most plausibly implicated in creating disparities. For example, in 2015, 3.8% of white 8th graders had a math teacher in their first year, compared with 6.1% of black students and 7.1% of Hispanic students (Rahman et al. 2017). Notably, differences across states are much larger than differences by social background; 0.2% of white students in Michigan had a novice teacher, compared with 9.7% in Hawaii. Comparably, 0.6% of black students in Alabama had a novice teacher, versus 13.1% in California.

While the conventional wisdom was previously that teachers gained little additional effectiveness beyond their first five years of experience, newer evidence on experience, however, suggests that there are returns to teacher experience across the career, particularly in math. Papay and Kraft (2015), using a variety of fixed effects models and data from a large urban school district, found that while growth in teacher contributions to test score gains was greatest in the first few years of teaching, there is also continuing evidence of returns to further years of experience. Papay and Kraft's results corroborate earlier findings from Jackson (2014), who uses data on 9th grade teachers in North Carolina and models teacher effects as the teacher's contribution to student outcomes net of lagged student achievement and a set of other covariates. Wiswall (2013) uses similar methods and data to study 5th grade teachers and reported similar dynamics: experience continued to contribute to growth in math scores beyond five years. Others have reported positive returns to experience beyond five years for English as well (Ladd and Sorenson 2012; Harris and Saas 2011).

Researchers have also found substantial effects of teacher experience on engagement-related outcomes. Ladd and Sorenson (2012) find that teacher experience has large effects on behavioral outcomes, particularly absenteeism; on average, teachers who obtain over 21 years of experience reduce levels of high student absenteeism by almost 15 percentage points relative to novice teachers. Jennings and DiPrete (2010) also find returns to kindergarten teacher experience on social and behavioral skills. Jackson (2014), however, studying high school teacher effects, found no evidence that experience was associated with better non-test score outcomes.

Though what exactly makes an effective teacher continues to be debated, studies examining the relationship between observable teacher characteristics and student outcomes have found that teacher experience plays an important role in teacher effectiveness. In the next section, we also consider the contribution of teachers' and students' demographic similarity to student engagement and achievement.

Teacher Race/Ethnicity and Gender Matches

To date, the majority of student-teacher race matching studies related to academic achievement have focused on black and white students and their teachers (McGrady and Reynolds 2012). These studies generally find positive effects of race matching on academic achievement. Dee (2004), analyzing data from the Tennessee STAR study, found evidence that both white and non-white students exposed to a teacher of their own race had higher academic achievement. Ehrenberg and Brewer (1995), re-analyzing data from the Coleman Report, found some evidence that in schools with a higher percentage of black teachers, white students experienced lower test gain scores and black students' gains increased, but this was not consistent across grade levels. Eddy and Easton-Brooks (2011), analyzing more recent data from the ECLS-K, found that black students taught by an African-American teacher in kindergarten had higher gains in math scores.

A small number of studies have examined race/ethnicity matching for Hispanic students. Three of these studies (Fraga, Meier, and England 1986; Meier 1993; Pitts 2007) only test the association of the proportion of Hispanic teachers and students with student performance at the district level. Pitts (2007) observed that when the proportion of black and Hispanic teachers and students in a Texas school district are a close match, a higher percentage of black and Hispanic students pass a standardized test required for graduation. Meier (1993) found that Hispanic students experienced a small but significant improvement in communications and math test passing, grade promotion, and lower dropout rates in districts with a higher percentage of Hispanic teachers. A similar pattern emerged from Fraga, Meier, and England (1986) in which a higher proportion of Hispanic students in a school district completed school and enrolled into college from high schools with a higher percentage of Hispanic teachers. Two additional studies (Clewell, Puma, and McKay 2005; Ehrenberg, Goldhaber, and Brewer 1995) examine the effect of racial and ethnic matching at the student level, but both do so with data that are two decades

old. Clewell, Puma, and McKay (2005) found that Hispanic fourth and sixth graders with same-ethnicity teachers had higher math test scores than Hispanic students with teachers of a different background. Hispanic fourth graders also improved in reading comprehension test scores. Ehrenberg, Goldhaber, and Brewer (1995) found that although a racial match between students and teachers was associated with changes in teacher perceptions of Hispanic, black, and white students, it did not appear to affect gains in test scores from eighth to tenth grade.

Another group of studies addresses the importance of teacher-student demographic matching on students' behavioral and emotional engagement. For example, Holt and Gersherson (2015) using fixed-effects models of student-level data in North Carolina, show that same-race teachers increase school attendance and decrease students' suspensions. Regarding the latter, Lindsay and Hart (2017) suggests that the likelihood of receiving different types of exclusionary discipline decreases if there is teacher-student racial congruence. The effects outlined by these scholars are small in magnitude, but highly robust across different specifications and grade spans, including the use of individual fixed-effects and instrumental variable examination.

Other studies, such as Egalite and Kisida (2016), show that race is not the only driver affecting students' behaviors and perceptions. These researchers conclude that there are large beneficial effects on perception of what could be called emotional engagement, the quality of the student-teacher relationship, and student motivation when the gender and racial make-up of teachers match those of the students. Similarly, Fox (2016) found a large effect of teacher racial matching for black students' expectation of completing college. The magnitude of having a same-race teacher for black students is over 70% of the White-Black racial gap in expectations.

A distinct pathway through which race matches appear to affect student achievement and engagement is through their recommendations for special program placement. Grissom and Redding (2016), conditioning on academic achievement, find that Black students are less likely to be referred to gifted programs when taught by non-black teachers. Along these lines, Fish (2017) using an original survey experiment, finds that teacher are more likely to perceive low academic performance as "expected" for black students, while see such performance among white students as a reason to seek out medical diagnoses. Consistent with these findings, Gershenson et al. (2015) show that non-black teachers have systematically lower expectations of black students' likely educational attainment. These findings suggest that teachers' judgement of

"exceptionality", in the context of special and gifted education, can be driven by subjective and racially biased constructs.

Lastly, Gershenson et al. (2017) using student-level data and an instrumental variable analysis approach, find that teacher-student demographic matches have a persistent effect over time. Assigning a same-race teacher between third and fifth grade, they find, significantly reduces the probability of high school dropout for black students. This effect is larger for economically disadvantaged black boys. As in other studies, this study posits that same-race teachers also increase the college aspirations of black students.

School Socioeconomic and Racial Composition

Since the publication of the Coleman Report (1966), research has documented considerable gaps in educational outcomes between students attending schools with high and low concentrations of poor and underrepresented minority students (Coleman et al. 1966; Mickelson, Bottia, and Lambert 2013; Palardy 2013; Rumberger and Palardy 2005; Wells and Crain 1994; Condron et al. 2013; Reardon and Owens 2014; Reardon 2016). However, because school composition is an outcome of choices made by both the state and families, it is difficult to isolate causal impacts of school and classroom composition. Nevertheless, evidence points to a relationship between the racial and SES composition of schools and the outcomes of students attending them.

Court desegregation orders provide some of the best quasi-experimental evidence about the effects of racial composition on academic achievement. The imposition of the court order provides a plausibly exogenous source of variation that allows researchers to identify an effect of a change in the racial composition on student outcomes. Johnson (2011) takes advantage of the timing in court orders, and finds that black students' educational attainment and graduation rates benefited from exposure to desegregation plans. Guryan (2004) uses Census data from the 70s and 80s and finds that the dropout rate for black students decreased by 2-3 percentage points after desegregation orders were imposed. Both Lutz (2011) and Saatcioglu (2010) use recent removals of desegregation orders to examine the effect of composition, and find that black students' likelihood of dropout increased after the dismissal of court-order desegregation plans. This is not consistent across all districts; Ludwig (2008), however, did not find that the white-black test score gap widened after the busing program ceased in Charlotte-Mecklenburg.

Other studies have focused on the persistence over time of the "equalizing" effect of desegregation plans. For instance, Ashenfelter et al. (2006) finds that effective desegregation had important long-term labor market implications for southern-born black adults. Using older birth-cohorts as a rough counterfactual, these authors find that black men's annual income increased between 5 to 9 percent. Johnson (2011), using a similar strategy, finds consistent and large long-run beneficial effects of court orders in terms of occupational attainment and adult earnings. Additionally, this study shows that effective desegregation had a positive impact on other relevant life outcomes, such as health and the probability of incarceration. Regarding the latter, Weiner et al. (2009) exploited variation across school districts to conclude that cohorts that attended desegregated schools experienced a persistent reduction in arrests. These benefits associated with desegregation appear to endure across generations: using multigenerational data from the Panel Study of Income Dynamics, Johnson (2012) found an increase in academic achievement, levels of educational attainment, and college quality/selectivity for the children of parents who attended schools under desegregation orders.

However, it is difficult to separate the effects of segregation from the effects of other factors, such as school funding, that are closely related to the racial/ethnic or income composition of the school. Reber (2011), for example, studied enrollment and graduation rates in Louisiana and found an increase in the graduation rate for black students as a consequence of desegregation plans, consistent with other studies on desegregation plans. However, rather than changes in the black-white school exposure, Reber suggests that changes in school funding explained these improvements.

Beyond studies using court orders as an analytic tool, two studies using data from Texas attempt to isolate the effect of composition by exploiting plausibly exogenous yearly variation in the demographic makeup of cohorts of students. As the makeup of the classes varied across years, the researchers could attribute differences in average test score performance to those changes. Hoxby (2000) found that exposure to a lower proportion of black students increases reading and math test scores for black students. Hanushek et al. (2009) found that attending school with a higher percentage of black schoolmates reduces achievement for black students. Neither of these studies examined the socioeconomic composition of schools.

Importantly, the reduction in White-Black disparities cited by these studies are driven mostly by improving outcomes of Black students, with no observable detrimental effects on their

White peers. Hoxby (2000) and Hanushek et al. (2009) noted the tiny or insignificant degree to which an increase in racial diversity influenced white students' achievement. Gamoran and An (2016) using school-by-grade fixed effects models to study the effect of a school desegregation case in Nashville on student achievement and found no evidence that an increase in diversity had a negative effect on achievement growth of white students. Additionally, Ackert (2018) conducted surveys of student engagement and reports an "affective-behavioral trade-off" for students in schools with high proportions of white students whereby students attending whiter schools are more likely to say that they are engaged in coursework, but less likely to report liking school.

While the evidence demonstrates relationships between school composition and achievement outcomes, recent work indicates that research may benefit from turning attention to settings that are becoming increasingly common: segregation for Latino students, as well as prekindergarten classrooms. Orfield, Kucsera, and Siegel-Hawley (2012) report dramatic increases in segregation for Latino students, who they find are "attending more intensely segregated and impoverished schools than they have for generations," but little work has been done on the effect of this increasing segregation. Reid and Ready (2013) study the association between socioeconomic classroom composition and children's social and cognitive development in prekindergarten classrooms. Using fixed effects models and data from the ECLS-K, they find positive relationships between the mean socioeconomic status of the class and children's receptive language, expressive language, and mathematics learning.

Some of the most convincing evidence on socioeconomic integration can be found in research by Schwartz (2010), who studied an instance in which students in public housing were randomly assigned to schools, thereby isolating the effect of schools on student outcomes. She finds that by their fifth year of elementary school, students from public housing in low-poverty elementary schools had significantly higher scores in math and reading than equally poor students assigned to high-poverty schools, and that those gaps grew over time. By the end of elementary school, children in public housing in Montgomery County's most affluent half of elementary schools performed two-fifths of a standard deviation higher in math and one fifth of a standard deviation higher in reading than otherwise similar children in public housing who attended schools with greater than 20 percent poverty. This growth made up for a substantial portion of the starting gap between the students in poverty and those not in poverty. While she

was unable to estimate a direct effect of composition per se, nor the specific mechanisms by which composition has an effect, Schwartz's work makes clear that the association between the composition of a student's school and their outcomes is not simply an artifact of selection effects.

Curricular Rigor and Gateway Courses

A longstanding body of educational research has argued that differences in exposure to challenging instruction plays a role in producing racial and socioeconomic outcome disparities (Gamoran 1987; Gamoran and Mare 1989; Oakes 1985). Early work identified inequality both within and between schools as sources of disparities (Gamoran 1987; Gamoran and Mare 1989). In other words, these differences can emerge because students of different backgrounds attend different schools, which in turn offer different instructional opportunities, or because students of different backgrounds are tracked into instructional settings within schools that are more or less rigorous. A key source of inequality is that of access – schools with higher percentages of poor and underrepresented minority students are less likely to offer higher level courses, placing students who attend those schools at a disadvantage.

Much of the research on expanding access to courses has focused on math, and in particular algebra, which serves as a "gateway" course for higher-level mathematics. Large racial disparities currently exist in students' likelihood of taking Algebra before high school; in 2009, 12% of black, 17% of Hispanic, 29% of white, and 48% of Asian students did so (Nord et al. 2011). Students who do not complete algebra in 8th grade struggle to finish a full course of mathematics in high school, which can become a formidable barrier to many careers (Attewell and Domina 2008; Long, Conger, and Iatorola 2012). Multiple districts and states have extended access to algebra to all 8th grade students, providing opportunities for researchers to study the effects of expanded coursetaking on student outcomes.

Disparities in gateway course completion exist for a multitude of reasons, including student, school, and family-level factors. More work is needed to understand all of the causes, but research on algebra and other programs has identified teacher subjectivity as a substantial cause of at least part of the gap. Grissom and Redding (2016) study the rates at which different types of students are referred to gifted programs and find that Black students are referred at significantly lower rates than white students with similar standardized test scores, even when

controlling for a range of other factors such as health, socioeconomic status, and classroom characteristics. They also find that similar black students taught by black teachers are more likely to be referred to gifted programs than those taught by non-black teachers. Thompson (2017), using a combination of quantitative and qualitative evidence, suggests that English Language Learners are held back by institutional, classroom, and individual factors, including, but not limited to, school or district-level course placement polices, self-motivation and prior achievement, and teacher expectations. Dougherty et al. (2015), using a regression discontinuity design and data from Wake County in North Carolina, found that assigning students to algebra based on a defined prior achievement metric reduced the relationship between course assignment and student characteristics such as income and race/ethnicity, indicating that discretionary barriers to the course may have held some students back before the policy change.

Evidence from expanded access to algebra courses demonstrate that there are significant gains to be made by increasing access to higher-level mathematics courses for qualified students and supporting them once they are there. Heppen et al. (2011) conducted a randomized control trial testing the impact of offering an online Algebra I course to students judged by their schools to be ready to take Algebra I in grade 8, but who attended schools where the course was not offered. They found that students with access to the online algebra course had higher levels of mathematics achievement and were more likely to enroll in higher-level mathematics courses in high school. In a follow-up paper to their work on assignment to algebra, Dougherty et al. (2017) employ their regression discontinuity framework to studying the effects of assignment to algebra. They find that applying an objective course assignment rule based on test scores identifies well-prepared students that more subjective systems might have overlooked, and that such students are likely to benefit from advanced placement, though benefits are concentrated among females and students who are not low-income. These benefits included increased likelihood of enrollment in higher-level courses, increased college readiness as measured by an exam, and a larger fraction of students indicating an intention to attend college. McEachin et al. (2017) also employed a regression discontinuity design and found that a policy change in California leading to more students enrolling in Algebra I classes had substantial positive effects on high school math course-taking and smaller, but still positive, effects on high school math and English achievement. Almost no work has been done on ELA or other courses.

On the other hand, research also demonstrates that there are negative achievement effects at the bottom of the distribution associated with pushing unprepared students into advanced courses (Penner et al., 2015; Domina et al. 2015; Clotfelter, Ladd and Vigdor 2015, Simzar, Domina, and Tran 2016, Loveless 2008). Studies that have found positive effects of expanded access typically employ research designs that identify qualified students or students close to the threshold for qualification. Research that has examined students on the lower end of the mathematics performance distribution has been much less positive. Clotfelter, Ladd and Vigdor (2015) take advantage of the timing of implementation of an accelerated algebra course in two North Carolina districts to study the effect of the course and find significant negative effects on performance in both algebra and geometry for previously lower achieving students. Simzar, Domina and Tran (2016) find that placing lower-performing students in higher level mathematics courses can undermine their motivation for achievement. However, in a randomized control trial, Cortes, Goodman, and Nomi (2015) find that low performing 9th grade students who received double-dose algebra exhibited substantially improved math test scores and probability of course completion, indicating that with extra support the negative effects of accelerated curricula might be attenuated.

In evaluating the impact of these courses, it is often difficult to establish the specific mechanisms associated with better outcomes, even as scholars have attempted to address non-random assignment to advanced coursework. As with the other factors identified in this section, it is unclear what mechanisms drive any direct effects of increased access to advanced coursework. For example, more challenging courses at the high school level typically enroll higher-ability peers and may also be allocated higher-quality teachers, but these mechanisms have remained largely untested. Instead, it may be more fruitful to consider how access to higher-level coursework creates the conditions under which greater equality is possible and achievable. It is clear that between-group inequality exists in access to gateway courses such as algebra, but that there is much that schools and districts can do to narrow that gap. While there is no one-size-fits-all solution, existing evidence suggests that attempts to expose more students to higher-level coursework and support them once they are there can be successful if they are carefully designed and implemented.

Attendance

In light of scholars' interest in understanding the school year factors that influence racial and income achievement gaps, multiple studies have considered the potential role of student absences in light of the large bodies of literature documenting the impact of time spent in school on student outcomes (Alexander et al. 2007), and the risk factors for poor student attendance that are highly associated with race/ethnicity and income. For example, lower-income and underrepresented minority students are more likely to be in poor health (Subramanian and Kennedy 2009), residentially mobile (DeLuca and Dayton 2009), exposed to violence and safety concerns in their neighborhoods (Sharkey 2010), charged with caring for younger or older family members (Lareau 2003), and less attached to school, particularly in the middle and high school years (Johnson, Crosnoe, and Elder 2001). Attendance is also important because it is a direct component of student engagement: analyses of the construct of engagement typically include attendance as one of multiple measures (Fredricks and McCloskey 2012; Appleton, Christenson, and Furlong 2008). Considered together, these bodies of literature suggest that school absences – both a measure of behavioral engagement and a predictor of future engagement - may affect income and racial/ethnic disparities in both academic achievement and engagement.

Nationwide data on absenteeism paints a clear and unequivocal picture of disparities by race/ethnicity, poverty, and disability status (Jacob and Lovett 2017; Gee 2018). In the 2013-2014 school year, black students were 50% more likely to be chronically absent than white students (defined as missing at least 15 days during the school year), at rates of 14.6% for Blacks compared to 9.7% for whites. Rates of chronic absenteeism are higher for students with disabilities (SWDs), and SWDs receiving more inclusive services less likely to be chronically absent (Gottfried et al. 2017).

Many studies have found a negative association between student absences and academic performance at all grade levels, from Head Start pre-k programs to high school (Ansari and Purtell 2017; Auecejo and Romano 2014; Benner and Wang 2013; Gershenson et al. 2015; Gottfried 2009, 2010, 2014; Ready 2018; Smerillo et al. 2018). These analyses vary by their treatment of absences, some treating them as continuous and others as a binary indicator of chronic absenteeism. Others have distinguished between types of absences (excused versus unexcused), and their relationship to school performance. Gottfried (2009), for example, used fixed effects modeling to determine that a higher proportion of unexcused absences in elementary school students was associated with academic risk, particularly in math, as compared

to students with a higher proportion of excused absences, which was positively associated with lower reading and math scores. While there is no experimental evidence on the causal effect of absences, Gottfried (2010) employs quasi-experimental methods, including instrumental variables and fixed effects, to estimate the effect of absences on GPA and test scores and finds a negative relationship. All of the available evidence suggests that absences have a negative effect on academic achievement.

There is also evidence suggesting that absenteeism influences other outcomes. Gottfried (2014) evaluates the effect of chronic absenteeism on achievement and socioemotional outcomes and finds that chronically absent students experience reductions in math and reading achievement outcomes as well as declines in social engagement. Gottfried (2015) considers the impact of chronic absenteeism on the absentee's classmates. He finds that after controlling for a student's own chronic absenteeism and other covariates, students with chronically absent classmates had test scores that were 0.04 standard deviations lower in reading and 0.05 standard deviations lower in math (Gottfried 2015).

Research has also found that the negative relationship between absenteeism and achievement increases as students move through school, with absences in later years having a greater negative impact than those in earlier years. Gershenson et al. (2015), for example, found larger test score reductions among fourth and fifth graders (0.007 and .004 test score standard deviation reductions in math and reading) compared to kindergarteners and first graders (0.002 test score standard deviation reductions).

Additional work has documented the relationships between socioeconomic status, school attendance, and academic achievement. Using random and within-child fixed effects models, Morrissey et al. (2013) finds that while poor attendance and low income both have negative impacts on students' grades, they appear to do so independently of each other. Ready (2018) also examines these relationships, reporting that socioeconomically disadvantaged children with good attendance rates gain more literacy skills than their higher-SES peers during kindergarten and first grade.

Many of these attendance-outcome relationships varied by subgroup. Ansari and Purtell (2017) find that Black and Latino children and children from households without two parents were less likely to be both absent and chronically absent from Head Start than white children or those who came from households with married parents. Children enrolled in larger or bilingual

classrooms, or classrooms that operated for more hours per week, were also less likely to be absent. Other recent work has found that girls and higher-ability students with more absences had larger negative associations between missing school and test scores than boys or lower-ability students (Gottfried and Kirskey 2017). In examining how school attendance patterns changed across middle and high school, Benner and Wang (2013) found that students from schools that were small, segregated, served larger percentages of economically disadvantaged students, and had less experienced teachers were more likely to shift to a lower attendance trajectory whereas students from large, diverse, and affluent schools were most likely to shift to a higher attendance trajectory. Smerillo et al. (2018) used propensity score methods on data from the Chicago Longitudinal Study and found that chronic absenteeism was associated with lower math achievement and a lower probability of completing high school.

There is evidence that student absenteeism can be addressed and improved. In some cases, this improvement is associated with interventions designed to increase family and community involvement with schools. Sheldon and Epstein (2004) examined chronic absenteeism rates at 39 elementary and secondary schools and found that school, family, and community partnership practices, including parent outreach, celebration of good attendance, and assignment of community mentors to chronically absent students, can significant decrease chronic absenteeism. A later study showed similar results, reporting that elementary schools that implemented these types of partnerships demonstrated a slight improvement in attendance compared to schools that did not develop such programs, whose attendance rates declined slightly year to year (Sheldon 2007).

Other studies have evaluated different kinds of outreach interventions. McCluskey et al. (2004) assessed a program designed to reduce truancy in three elementary schools, which targeted students with chronic attendance problems in a multi-step intervention that began by sending home a letter from the principal informing parents of the number of school days their child had missed and, if attendance remained unimproved, could involve referral to an attendance officer, a community mental health agency, a child and family social service agency, and, eventually, a community-policing officer. The authors found that attendance significantly improved after the first two stages of the intervention (principal letter and attendance officer visit), and reported than only 20% of participants required further intervention than the initial principal letter. DeSocio et al. (2010) also piloted a truancy intervention project in an urban high

school with promising results. Students with more than 15 unexcused absences were randomly assigned to an intervention group, which included enrollment in a school-based health center and assignment of a teacher mentor, or a control group, which received only regular school services, and students in the intervention group were significantly more likely to remain in school than students in the control group.

Other studies have focused on interventions that rely on interagency collaboration, in which schools work with other organizations, including judicial systems, law enforcement agencies, and service providers (Fantuzzo et al. 2005, Richtman 2007). In a review of Project START (Stop Truancy and Recommend Treatment), a program designed to reduce truancy through community-based interventions, authors found that students referred to multidimensional, community-based family court (Project START) had a significant drop in their rates of unexcused absences following the intervention, and maintained these reduced truancy rates at 30 and 60 days and one year post-intervention (Fantuzzo et al. 2005). Similarly, the Attorney General in Ramsay County, which serves the Minneapolis-St. Paul area, developed the Truancy Intervention Program to reduce truancy, increase school connectedness, and improve high school graduation rates (Richtman 2007). Following the implementation of this program, in which students with poor attendance were identified and provided with progressively intrusive interventions starting with a meeting, followed by an attendance hearing, and the eventual filing of a truancy petition, the number of students missing 15 or more days of schools decreased by more than 50 percent (Richtman 2007).

Availability of and Participation in Extracurricular Activities

Extracurricular activities can serve multiple functions. They can provide opportunities to make connections and develop prosocial ties, fostering a sense of inclusion in a community that at once helps to keep students engaged and enrolled in schools while at the same time socializing students into a set of norms and values that will help them be successful later in life.

Very little work has been done to understand differences between groups in terms of access to and participation in extracurricular activities. One exception is Stearns and Glennie (2010), who painstakingly matched activities lists from high school yearbooks with school-level data in North Carolina and found that school size and poverty levels are significantly related to

both the number and types of activities that schools offer. Larger schools and schools with a lower proportion of students in poverty were more likely to offer more and more varied types of activities. Additionally, the number of opportunities is associated with higher participation rates.

Holland and Andre (1987) reviewed 20 early years of research on extracurricular involvement and find that associations between participation and adolescent functioning are generally small but positive. This broad conclusion was supported by later reviews encompassing more recent research (Feldman and Matjasko 2005; Farb and Matjasko 2012). Shulruf (2010) also reviews research on different types of extracurricular activities and finds a variety of small to moderate relationships between participation and academic performance in Math and English.

Students do not randomly select into extracurricular activities, and to some extent any observed effects are due as much to the quality of the match between student and activity as to the fact of participation in an activity. Causal estimates of effects are typically calculated using fixed effects and instrumental variables approaches using longitudinal nationally representative surveys, but these estimates still suffer from selection problems and issues of generalization. Rees and Sabia (2010) use data from the National Longitudinal Study of Adolescent Health and employ two methods, fixed effects models and height as an instrumental variable, to predict the effect of participation in sports on GPA, aspirations to attend college, difficulty paying attention in class, and difficulty completing homework. They find small positive effects on GPA, aspirations, and difficulty completing homework. Lipscomb (2007) uses National Educational Longitudinal Study (NELS) data and fixed effects models, and finds small positive effects on test scores for participation in high school sports.

The mechanisms through which extracurricular activities influence student outcomes include engagement in school, aspirations, social capital, identity formation, peer group membership, and attachment to non-familial adults (Broh 2002; Eccles and Barber 1999; Eccles et al. 2003). For the most part, research on mechanisms is correlational. Broh (2002), for example, using OLS regressions and data from NELS 1988, found the strongest relationships with academic achievement for sports and student government, which she attributed to those activities' power to promote students' development and social ties. Fredricks and Eccles (2006) collected longitudinal survey data on 1500 students in Maryland starting in 7th grade and continuing past high school. They found that participation in 11th grade school clubs and

prosocial activities was associated with continuing schooling and civic engagement one year after high school, but higher sample attrition among African Americans and at-risk youth raises questions about the generalizability of results. Im et al. (2016) used propensity score weighting on a sample of 465 students to study the effects of extracurricular participation on motivation, valuing of education, teacher-rated engagement, and competence beliefs. They found that participation in sports predicted competence beliefs and valuing education while participation in performance arts/clubs predicted teacher-rated engagement and grades.

The available evidence suggests that extracurricular activities have positive effects on student engagement and academic performance, but a lack of administrative data on the availability of and participation in activities severely limits what we know. The studies reviewed attempted to estimate the direct effects of participation in extracurricular activities, which are but one example of how disparities could arise; other enabling effects might also be equally or more important. The presence or absence of extracurriculars is indicative of opportunities available to students in a school, and the evidence available suggests that those opportunities are unequally distributed across schools. If that unequal distribution also influences the distribution of engagement and aspirations, simply measuring the direct effects of participation would underestimate the between-school or between-group disparities. Additionally, patterns of participation in activities may vary within schools as well, creating further possibilities for stratification.

IV. Proposed Indicators

Our choice of proposed indicators is driven by 3 principles. **The first is extreme parsimony**. While earlier indicator projects attempted to describe the health of multiple parts of the system, competition for attention is substantial (VanKnippenberg et al. 2015) and too much information can impede decision-making. **The second is local salience**. Education is fundamentally a state and local function. If indicators are intended to initiate policy change, they need to generate action at these levels. We believe choosing measures that allow disaggregation to state/local levels is the most effective ways to do this. **The third principle is to capture inputs as we believe they affect students over their educational life courses by measuring cumulative exposure**. Most existing indicators are limited by reporting exposure only in one

grade-level or time period. This approach, we believe, understates the magnitude of the differences between groups. While there are certainly data challenges created by this proposal, we believe that they are tractable given the widespread implementation of state longitudinal data warehouses.

Measures of cumulative exposure differ from those of cross-sectional exposure for both mechanistic and theoretical reasons. Mechanistically, if exposure at any one point in time causes a change in an outcome, then within a given range of possible effects cumulative exposure will cause cumulative change. In the real world, effects are unlikely to be quite so mechanistic. Instead, cumulative exposure to risks is both harmful on its own and indicative of an individual's social position. Theoretically, this implies that understanding the effects of an input requires an in-depth theory about how that input operates to influence outcomes, how selection into different lengths of exposure might vary, and how those two relationships interact over time: what Wodtke, Harding, and Elwert (2011) call a "complex time-dependent process of selection, exposure and feedback" (p. 731). Models that do not take these factors and relationships into account can understate the magnitude of effects because by controlling for interrelated variables they "control away" part of the effect itself. For example, while Harding (2003) uses propensity score matching estimates from the Panel Study of Income Dynamics and finds that exposure to high-poverty neighborhoods (greater than 20%) compared to low-poverty (less than 10%) during adolescence reduces the odds of high school graduation by about 50 percent for blacks and non-blacks alike, Wodtke, Harding, and Elwert (2011) use the same data and time-dependent methods and find that sustained exposure to disadvantaged neighborhoods is associated with decreases in the odds of high school graduation of about 80 percent for blacks and 60 percent for nonblacks. Given that blacks are more likely to live in high-poverty neighborhoods, and to stay there longer, these disparities in estimates are evidence that cross-sectional analyses understate effects of exposure to risk.

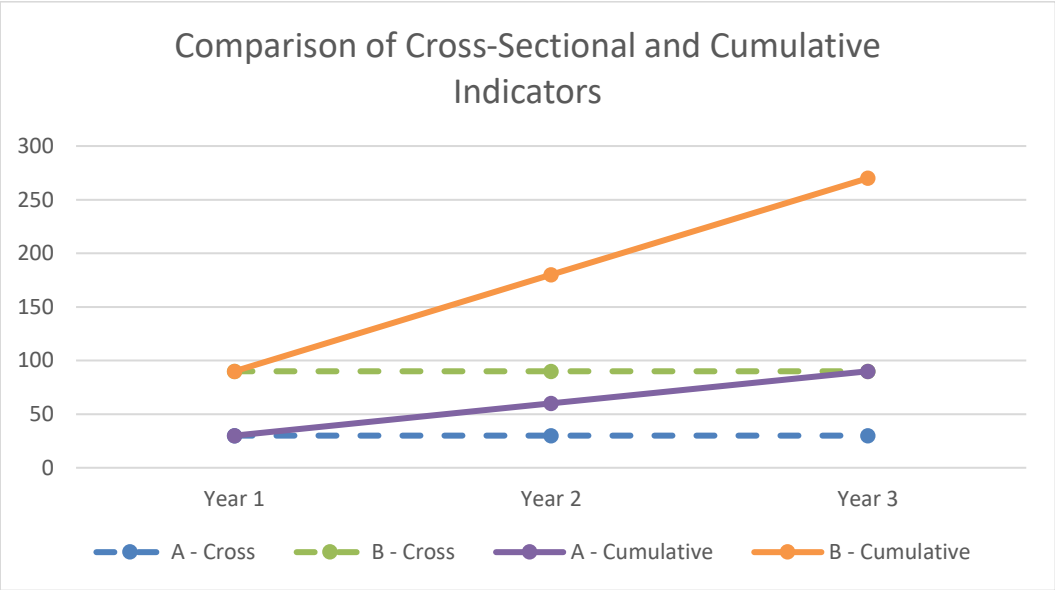
Some of the most powerful evidence for the importance of measuring cumulative exposure rather than cross sectional come from the study of neighborhood effects. Sharkey and Faber (2014) contend that it is important to consider the effect of contextual variables, like neighborhoods, within a life-course framework that tracks the emergence and growth of inequalities over time. Research has found effects of cumulative exposure over and beyond cross-sectional exposure to neighborhoods on a number of outcomes, including nonmarital

fertility (South and Crowder 2010), health (Clarke et al. 2014), academic achievement and behavior (Anderson, Leventhal, and Dupéré 2014), and high school graduation (Wodtke, Harding, and Elwert 2011). Estimates of the effects of cumulative exposure are typically much larger than cross-sectional estimates, implying that cross-sectional estimates severely understate relationships. Sharkey and Elwert (2011) extend their analysis beyond one generation and find effects of multigenerational exposure to neighborhood poverty on child cognitive ability. Little attention has been paid to effects of school-level cumulative exposures (though see Halpern-Manners 2009, 2016), and the important findings from the neighborhood effects literature underscore that gap in our understanding.

We propose an ambitious, large-scale effort to turn towards cumulative instead of cross-sectional indicators. In the interest of parsimony and simplicity of use, we advocate for arithmetically cumulative indicators. While not immediately intuitive, they are more accurate reflections of cumulative exposure to contexts than a running average because they reflect the widening of gaps over time in a way that averages do not. For example, consider two students, A and B, in different schools. Student A's school has a concentration of 30% students who received free and reduced-price lunch, while Student B's school served a more disadvantaged population and 90% of student receive free and reduced price lunch. In the first year of their schooling, the difference between A and B's concentration of free and reduced-price lunch would be 60 percentage points. If they remained in their same schools each year and those schools' populations did not change, then the difference between student A's and student B's context would remain at 60 percentage points throughout all three years and would fail to capture the fact that remaining in an environment of concentrated disadvantage over multiple years is qualitatively different than doing so for one year. A measure that tracks cumulative exposure would demonstrate that the gaps between the two actually grows from year to year.

Figure 1 illustrates the difference between the two indicators for student A and student B. While using a cross sectional indicator would imply that the difference between the two students remains constant throughout their educational trajectories, a cumulative indicator clearly demonstrates that the gap widens as years pass. Importantly, a running average or cumulative mean across years also understates the growth of the gap in the same way as a cross-sectional indicator does.

Figure 1: Comparison of Indicator Trajectories



1) Indicators for Principal Quality: Principal and Teacher Turnover

The research reviewed above demonstrates that exposure to high quality principals and teachers is positively associated with student achievement outcomes. We propose that one measure of between-group equity is average exposure to teacher and principal turnover. Turnover has direct effects on students through the disruption of the school community, but it is also a diagnostic indicator that points to differences in other important variables like the concentration of disadvantage and the likelihood that high-quality teachers and principals will remain in a school.

The most common measure of principal turnover is a binary variable (i.e. was the principal new to the school, or not), which could be summed over all years of a students' education, and within cohorts across students who share a subgroup characteristic. Similarly, teacher turnover could be measured as the number of teachers who replaced departed teachers, divided by the total number of teacher positions in the previous year, summed over students' educational careers.

These measures would have the advantage of being a proxy for teacher satisfaction with their job relative to other potential schools or even occupations, indicating the extent to which teachers or principals wish to remain at the school, without relying on subjective reports of working conditions. Research has found strong negative relationships between school working conditions and teacher turnover (Ladd 2011), as well as substantially higher turnover in schools with large fractions of racial minorities or students from low-income families (Clotfelter, Ladd, and Vigdor 2010; Jackson, 2009; Hanushek, Kain, and Rivkin 2004). As rates of movement are generally higher to more white and affluent schools for principals (Cullen and Mazzeo 2007; Branch et al. 2008; Clark et al. 2009; Miller 2009; Gates et al. 2006; Horng, Kalogrides and Loeb 2009) and teachers (Jackson 2009; Clotfelter, Ladd, and Vigdor 2010), even to the extent that the race and achievement of students is more predictive than salary differences (Hanushek, Kain, and Rivkin 2004), turnover itself is a meaningful indicator that currently exhibits marked between-group variation and is significantly related to principal quality (Loeb 2011; Supovitz, Sirinides, and May 2010; Burkhauser 2017; Boyd et al. 2011; Grissom 2011).

2) *Indicators for Teacher Quality: Exposure to Novice and Experienced Teachers & Racial/Ethnic Teacher Matches*

While it is clear that teacher quality is important for student outcomes, it is less clear what indicator should be used to measure it. Some potential candidates, such as test-score based value-added measures, relate solely to one dimension of teacher quality. As is well-known, a focus on narrow indicators can sometimes produce behaviors that invalidate the nature of the inferences we want to make from that indicator to the domain it represents (Koretz 2017).

We propose an indicator that is both difficult to manipulate and related to a broad swathe of other measures of teacher quality: student exposure to experienced teachers. Specifically, we propose two indicators. First, we suggest that an indicator capture the cumulative percentage of students' teachers that are novices. That is, the sum of the ratios of novice teachers to non-novice teachers in each year. Second, we propose an indicator that is the simple average of the years of experience for all teachers a student regularly sees each year, summed over all years of a student's education. Each indicator could be averaged at the subgroup level within cohorts.

Cumulative exposure to teacher experience is a diagnostic indicator because teacher experience has both direct and enabling effects while also being indicative of a broader set of conditions about the school in which they work. Research reviewed above demonstrates that teacher experience has substantial direct effects on test scores (Papay and Kraft 2015; Jackson 2014; Wiswall 2013; Ladd and Sorenson 2012; Harris and Saas 2011) and non-cognitive outcomes (Ladd and Sorenson 2012; Jennings and DiPrete 2010). More experienced teachers are more likely to have had time to be mobile and have their choice of school or even occupation. National data show that disparities exist in the average levels of experience of teachers in schools serving more and less advantaged populations, indicating that inequalities in the distribution of experienced teachers is potentially demonstrative of other types of inequality, such as teacher working conditions.

Such an indicator would be clear, concise, and difficult to manipulate. If, for example, policymakers offered incentives for experienced teachers to move to or remain in schools serving disadvantaged students, those incentives could achieve their desired effect without drastic behavioral downsides.

Research has shown that matching student and teacher characteristics can have important effects on student outcomes. This result holds for matches on a variety of outcomes, including test scores (McGrady and Reynolds 2012; Dee 2004; Ehrenberg and Brewer 1995; Eddy and Easton-Brooks 2011; Fraga, Meier, and England 1986; Meier 1993; Pitts 2007; Clewell, Puma, and McKay 2005; Ehrenberg, Goldhaber, and Brewer 1995), expectations (Fish 2017), aspirations (Gershenson et al. 2015) and suspensions (Holt and Gershenson 2015; Lindsay and Hart 2017), and it persists over time (Gershenson et al. 2017).

We propose a diagnostic indicator that tracks the cumulative level of teacher-student demographic matches. Specifically, the measure would be calculated as the sum of the fraction of a student's teachers with whom they share the race/ethnicity or gender. This value could then be averaged across all students in a subgroup for each cohort. Additionally, one could track this teacher-student match indicator for specific subjects. This indicator would be especially helpful for tracking the relationship between high school coursetaking and pathways for specific subgroups of students into specific careers or college majors. For example, researchers continue to search for an explanation for the early emergence of a gender gap in college students majoring in STEM subjects. Having ruled out systematic variation in achievement (Riegle-Crumb et al.

2012; Justman and Méndez 2018), recent work has turned to variables like differential teacher treatment of students in STEM classes based on gender (McKellar et al. 2018). The micro-social interactions between teachers and students create the environment in which student develop or are deterred from pursuit of academic interests (whether specific or general), and a measure of the extent to which students see adults like them in positions of authority within that field is an important indicator.

Additionally, such an indicator would be difficult to manipulate without altering the actual conditions on the ground that have direct and enabling effects on student outcomes. The supply of types of teachers in specific classrooms is a function of the supply of teachers in general and the likelihood that a teacher will teach a particular class. Altering either of those variables to achieve a greater student-teacher match would imply changing a broad set of conditions that are indicative of the health of the teacher supply pipeline, the distribution of teachers to schools, and the distribution of teachers to subjects within schools. Thus, a change in the indicator of cumulative student-teacher demographic match would be a useful piece of information for policymakers, researchers, and practitioners for evaluative, descriptive, and predictive purposes.

3) Indicators for School Composition: Exposure to Poverty and Racial/Ethnic Groups

The available research demonstrates that exposure to concentrated poverty and segregation are related to student outcomes. Evidence from court desegregation orders (Johnson 2011; Guryan 2004; Lutz 2011; Saaticioglu 2010) and quasi-experiments (Schwartz 2010; Hoxby 2000; Hanushek et al. 2009) illustrates that changes in the socioeconomic composition of a student's school can influence their academic achievement, with potential effects later in life (Johnson 2012; Johnson 2011; Ashenfelter et al. 2006; Weiner et al. 2011). As well as being closely related to student outcomes, differential subgroup rates of exposure to poverty and racial/ethnic diversity are indicative of exposure to a range of conditions related to equality of access to opportunity within educational systems. They are also parsimonious, easy to interpret and communicate, and difficult to manipulate without changing underlying conditions.

As a measure of students' cumulative exposure to poverty within their school, we propose an arithmetically cumulative indicator that tracks the percentage of a student's classmates who receive free or reduced-price lunch. We also propose an identical measure that

tracks the racial/ethnic composition of a student's school. This measure would be composed of the percentage of students in a student's school who identify as members of a particular racial/ethnic group. Thus, one student would have an indicator for cumulative exposure to each racial/ethnic group and to poverty, each of which could be averaged across students in a subgroup to create subgroup-level indicators.

4) Indicator for Curricular Rigor and Gateway Courses: Completion of Course Sequences

Constructing an indicator for completion of gateway courses is quite difficult if one wishes to avoid indicators with the potential to be manipulated. The "algebra for all" effort in California provides some guidance. Domina and McEachin (2016) studied the ways in which middle schools responded to these pressures. They found that most commonly, schools either "detracked" by essentially forcing all eighth-grade students to take algebra, eliminating the barriers between what used to be tracks, or they "tracked up" by creating high-level course tracks for advanced students while enrolling lower-performing students in algebra. While the first should have equalizing effects, the second would preserve the status quo of separate tracks for different types of students.

We believe that a good indicator of equity is the extent to which subgroups complete *sequences* of courses. One potential way to do this is to follow Domina and Saldana (2012), who study cohort trends in enrollment in the High School and Beyond data. Enrollment in the next course in a sequence (e.g. calculus after trigonometry) indicates completion of the previous course. We propose keeping track of subgroup averages of enrollment levels in a series of courses, starting in the middle grades (pre-algebra) and continuing through calculus in high school, as an indicator of academic progression. This indicator would allow us to measure between-group differences in the proportion of students on track to complete a full sequence of courses, and to identify where in a course progression gaps emerge or widen.

A comprehensive, longitudinal indicator of course enrollment would be a positive step toward measuring exposure to curricula. Because it would track students or subgroups over time through multiple contexts, it would be difficult for individual schools to alter without affecting student preparation for the next course in the sequence. One worry might be that schools or systems would lower course entry requirements in order to permit more students to enroll, while "tracking up" and creating higher level courses to maintain inequality. However, the research

reviewed above indicates that on the margin more students can be enrolled in higher-level coursework than they currently are.

5) *Indicators for Processes Associated with Engagement: Attendance Rates and Opportunities for Participation in Extracurricular Activities*

Attendance is both an indicator of engagement (Fredricks and McColskey 2012; Appleton, Christenson, and Furlong 2008) and a predictor of future levels of engagement and other attainment outcomes (Gottfried 2010, 2014; Gershenson et al. 2015). Yet research on absenteeism is largely confined to cross-sectional work in which the outcome of interest is the relationship between rates of attendance in a given year and some metric. A cumulative exposure approach to studying absenteeism would take into account the time-dependent nature of progress through school as well as the feedback effects that various variables have on each other. We propose tracking cumulative rates of chronic absenteeism: the number of years in which a student was chronically absent. These values could then be averaged over all students in a subgroup to create a subgroup-level indicator.

We have very little information on both the availability of extracurricular activities and rates of student involvement in the activities that are available. Most studies of the availability of extracurricular activities are based on nationally representative surveys (e.g. NELS; Ad Health), though some employ different levels of data, such as matched yearbooks (Stearns and Glennie 2010) or smaller-scale survey data (e.g. Fredricks and Eccles 2006). A first step toward tracking opportunities for participation in extracurricular activities would be to collect the number of activities available in schools. These could be averaged across all students who are members of a subgroup to create subgroup-level indicators. Existing evidence suggests that there are wide gaps between racial/ethnic subgroups in North Carolina (Stearns and Glennie 2010), but we simply do not have reliable statistics on a nationwide scale.

National surveys do provide information on student participation in extracurricular activities. We propose an indicator that tracks the number of extracurricular activities in which students are involved, summed over all years of their education.

While we extensively reviewed the literature to identify inputs and processes that are consistently associated with disparities, that context is central in the production of achievement and engagement means that any researcher's recommendations must be submitted with humility about our incomplete understanding of how inputs function in different settings and for different types of students.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Ackert, E. (2018). Segregation paradox? School racial/ethnic and socioeconomic composition and racial/ethnic differences in engagement. *Social Science Research, 70*, 144-162.

Adelman, C. (2006). The toolbox revisited: Paths to degree completion from high school through college. Washington, DC: US Department of Education.

Agodini, R., Harris, B., Thomas, M., Murphy, R., Gallagher, L., & Pendleton, A. (2011). Achievement effects of four early elementary school math curricula: Findings for first and second graders. *National Center for Education Evaluation and Regional Assistance*.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review, 72*(2), 167-180.

Allensworth, E. (2013). The use of ninth-grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk (JESPAR)*, *18*(1), 68-83.

Allensworth, E., Nagaoka, J., and Johnson, D.W. (2018) High school graduation and college readiness indicator systems: What we know, what we need to know. Chicago, IL: University of Chicago Consortium on School Research.

Anderson, D. M., & Walker, M. B. (2015). Does shortening the school week impact student performance? Evidence from the four-day school week. *Education Finance and Policy*, *10*(3), 314-349.

Ansari, A., & Purtell, K. M. (2017). Absenteeism in Head Start and children's academic learning. *Child Development*, *0*(0), 1–11.

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 153–178.

Ashenfelter, O., Collins, W. J., & Yoon, A. (2006). Evaluating the role of Brown v. Board of Education in school equalization, desegregation, and the income of African Americans. *American Law and Economics Review*, *8*(2), 213-248.

Attewell, P. A., Lavin, D. E., Domina, T., & Levey, T. (2006). New evidence on college remediation. *The Journal of Higher Education*, *77*(5), 886–924.

Attewell, P., & Domina, T. (2008). Raising the bar: Curricular intensity and academic performance. *Educational Evaluation and Policy Analysis*, *30*(1), 51–71.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279-1333.

Avery, C., Gurantz, O., Hurwitz, M., & Smith, J. (2016). Shifting college majors in response to advanced placement exam scores. *NBER Working Paper Series*, 56.

Béteille, T., Kalogrides, D., & Loeb, S. (2011). Stepping stones: Principal career paths and school outcomes. *NBER Working Paper No. 17243*, *53*(9), 1689–1699.

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, *20*(3), 899-911.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647-663.

Blase, J., & Blase, J. (1999). Principals' instructional leadership and teacher development: Teachers' perspectives. *Educational Administration Quarterly*, *35*(3), 349–378.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 1–25.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73(2), 125–230.

Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America* (Vol. 57). New York, NY: Basic Books.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, *48*(2), 303–333.

Branch, G. F., Hanushek, E. a., & Rivkin, S. G. (2012). Estimating the effect of leaders on public sector productivity: The case of school principals. *NBER Working Paper*, *17803*(January), 1–50.

Brewer, D. J. (1993). Principals and student outcomes: Evidence from U.S. high schools. *Economics of Education Review*, *12*(4), 281–292.

Brighouse, H. (2006). *On education*. Psychology Press.

Broh, B. A. (2002). Linking extracurricular programming to academic achievement: Who benefits and why? *Sociology of Education, 75*(1), 69.

Bryk, A. S., & Hermanson, K. L. (1993). Chapter 10: Educational indicator systems: Observations on their structure, interpretation, and use. *Review of Research in Education, 19*(1), 451-484.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Burkhauser, S. (2017). How much do school principals matter when it comes to teacher working conditions? *Educational Evaluation and Policy Analysis, 39*(1), 126–145.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and program planning, 2*(1), 67-90.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. *NBER Working Paper No. 17699*.

Clark, D., Martorell, P., & Rockoff, J. (2009). School principals and school performance. Working Paper 38. *National Center for Analysis of longitudinal data in Education research*.

Clewell, B. C., Puma, M. J., & McKay, S. A. (2001). Does it matter if my teacher looks like me?: The impact of teacher race and ethnicity on student academic achievement. *Urban Institute, Education Policy Center.*

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2011). Teacher mobility, school segregation, and pay-based policies to level the playing field. *Education Finance and Policy, 6*(3), 399-438.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2015). The aftermath of accelerating algebra: Evidence from district policy initiatives. *Journal of Human Resources, 50*(1), 159–188.

Cobb, C. W., & Rixford, C. (1998). *Lessons learned from the history of social indicators* (Vol. 1). San Francisco: Redefining Progress.

Coelli, M., & Green, D. A. (2012). Leadership effects: School principals and student outcomes. *Economics of Education Review, 31*(1), 92–109.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., and York, R.L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S.

Government Printing Office.

Condron, D. J. (2009). Social class, school and non-school environments, and black/white inequalities in children's learning. *American Sociological Review, 74*(5), 685-708.

Condron, D. J., Tope, D., Steidl, C. R., & Freeman, K. J. (2013). Racial segregation and the black/white achievement gap, 1992 to 2009. *The Sociological Quarterly*, *54*(1), 130-157.

Constance A.L., C.M.D. Hart. (2017). Exposure to same-race teachers and student disciplinary outcomes for black students in North Carolina. *Educational Evaluation and Policy Analysis*. *39*(3):485–510.

Corcoran, S. P., & Evans, W. N. (2008). Stalled progress in closing the race achievement gap: The Role of teacher quality. *Steady Gains and Stalled Progress: Inequality and the Black-White Test Score Gap, Russell Sage Foundation Press.*

Corcoran, S. P., Schwartz, A. E., & Weinstein, M. (2012). Training your own: The impact of New York City's aspiring principals program on student achievement. *Educational Evaluation and Policy Analysis*, *34*(2), 232–253.

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra. *Journal of Human Resources*, *50*(1), 108–158.

Cullen, J., & Mazzeo, M. (2008). Implicit performance awards: An empirical analysis of the labor market for public school administrators. *University of California, San Diego,* (December). Retrieved from http://www.kellogg.northwestern.edu/faculty/mazzeo/htm/txppals_1207.pdf

Darling-Hammond L. (1992). Educational indicators and enlightened policy. *Educational Policy* 6(3), 235-265.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195-210.

DeLuca, S., & Dayton, E. (2009). Switching social contexts: The effects of housing mobility and school choice programs on youth outcomes. *Annual Review of Sociology, 35*, 457-491.

DeSocio, J., VanCura, M., Nelson, L.A. Hewitt, G., Kitzman H., & Cole, R. (2010). Engaging truant adolescents: Results from a multifaceted intervention pilot. *Preventing School Failure: Alternative Education for Children and Youth, 53*(3), 3-9.

Dhuey, E., & Smith, J. (2014). How important are school principals in the production of student achievement? *Canadian Journal of Economics*, *47*(2), 634–663.

DiPrete, T.A. & Buchmann, C. (2006). Gender-specific trends in the value of education and the emerging gender gap in college completion. *Demography*, *43*(1), 1-24.

Domina, T. (2014). The link between middle school mathematics course placement and achievement. *Child Development*, *85*(5), 1948–1964.

Domina, T., & Saldana, J. (2012). Does raising the bar level the playing field?: Mathematics curricular intensification and inequality in American high schools, 1982-2004. *American Educational Research Journal*, *49*(4), 685–708.

Domina, T., Hanselman, P., Hwang, N., & McEachin, A. (2016). Detracking and tracking up: Mathematics course placements in California middle schools, 2003-2013. *American Educational Research Journal* (Vol. 53).

Domina, T., McEachin, A., Penner, A., & Penner, E. (2015). Aiming high and falling short: California's eighth-grade algebra-for-all effort. *Educational Evaluation and Policy Analysis*, *37*(3), 275–295.

Dougherty, S. M., Goodman, J. S., Hill, D. V., Litke, E. G., & Page, L. C. (2015). Middle school math acceleration and equitable access to eighth-grade algebra: Evidence from the Wake County Public School System. *Educational Evaluation and Policy Analysis*, *37*(1 Suppl), 80S–101S.

Dougherty, S. M., Goodman, J. S., Hill, D. V., Litke, E. G., & Page, L. C. (2017). Objective course placement and college readiness: Evidence from targeted middle school math acceleration. *Economics of Education Review*, *58*, 141–161.

Downey, D. B., Von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, *69*(5), 613-635.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of educational psychology*, *104*(2), 439.

Duckworth, A. L., & Seligman, M. E. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939-944.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... &

Sexton, H. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428.

Duncan, O. D. (1974). Developing social indicators. *Proceedings of the National Academy of Sciences*, *71*(12), 5096-5102.

Eberts, R. W., & Stone, J. A. (1988). Student achievement in public schools: Do principals make a difference? *Economics of Education Review*, *7*(3), 291–299.

Eccles, J. S., & Barber, B. L. (1999). Student council, volunteering, basketball, or marching band: What kind of extracurricular involvement matters? *Journal of Adolescent Research*, *14*(1), 10–43.

Eccles, J. S., Barber, B. L., Stone, M., & Hunt, J. (2003). Extracurricular activities and adolescent development. *Journal of Social Issues*, *59*(4), 865–889.

Eddy, C.M., & Easton-Brooks, D. (2011). Ethnic matching, school placement, and mathematics achievement of African American students from kindergarten through fifth grade. *Urban Education 46*(6), 1280-1299.

Egalite, A.J. & Kisida, B. (2016). The effects of teacher match on students' academic perceptions and attitudes. *Educational Evaluation and Policy Analysis*, *40*(1), 59-81.

Ehrenberg, R. G., & Brewer, D. J. (1995). Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. *Economics of Education Review*, *14*(1), 1-21.

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *ILR Review*, *48*(3), 547-561.

Estrada, P. (2014). English learner curricular streams in four middle schools: Triage in the trenches. *Urban Review*, *46*(4), 535–573.

Fantuzzo, J., Grim, S. & Hazan, H. (2005). "Project START: An evaluation of a community-wide school-based intervention to reduce truancy." *Psychology in the Schools, 42*(6), 657-667.

Farb, A. F., & Matjasko, J. L. (2012). Recent advances in research on school-based extracurricular activities and adolescent development. *Developmental Review, 32*(1), 1–48.

Farkas, G. (2003). Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology*, *29*(1), 541-562.

Feldman, F, A., & Matjasko, L, J. (2005). The role of school-based extracurricular activities in adolescent development: A comprehensive review and future directions. *Review of Educational Research, 75*(2), 159–210.

Fischer, S., & Argyle, D. (2018) Juvenile crime and the four-day school week. *Economics of Education Review, 64*, 31–39

Fish, E. (2017). The racialized construction of exceptionality: Experimental evidence of race/ethnicity effects on teachers' interventions. *Social Science Research* 62:317-334

Fox, L. (2016). Seeing potential: The effects of student–teacher demographic congruence on teacher expectations and recommendations. *AERA open*, *2*(1), 2332858415623758.

Fraga, L. R., Meier, K. J., & England, R. E. (1986). Hispanic Americans and educational policy: Limits to equal access. *The Journal of Politics*, *48*(4), 850-876.

Fredricks, J. A., & Eccles, J. S. (2006). Is extracurricular participation associated with beneficial outcomes? Concurrent and longitudinal relations. *Developmental Psychology*, *42*(4), 698–713.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement*, (pp. 763–782). Springer Science & Business Media.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59-109.

Fryer Jr, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447-464.

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education*, *60*(3), 135–155.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality : Compensation, reinforcement, or neutrality ? *American Journal of Sociology*, *94*(5), 1146–1183.

Gamoran, A., & An, B. P. (2016). Effects of school segregation and school resources in a changing policy context. *Educational Evaluation and Policy Analysis*, *38*(1), 43-64.

Gee, K. A. (2018). Minding the gaps in absenteeism: Disparities in absenteeism by race/ethnicity, poverty and disability. *Journal of Education for Students Placed at Risk, 23*(1–2), 204–208.

Gershenson, S., Jacknowitz, A., & Brannegan, A. (2017). Are student absences worth the worry in U.S. primary schools? *Education Finance and Policy, 12*(2), 137–165.

Gershenson, S., Hart, C.M.D., Lindsay, C.A., Papageorge, N.W. (2017). The long-run impacts of same-race teachers. IZA Discussion Papers, No. 10630 Working Paper.

Gershenson, S., Holt, S.B., Papageorge, N.W. (2015). Who believes in me? The effect of student-teacher demographic match on teacher expectations. IZA Discussion Papers, No. 9202.

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). From testing to teaching: The use of interim assessments in classroom instruction. Consortium for Policy Research in Education. Retrieved from http://repository.upenn.edu/cpre_researchreports/58

Goldring, R., and Taie, S. (2014). Principal attrition and mobility: Results from the 2012–13 Principal Follow-up Survey. First Look. NCES 2014-064. U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [date] from http://nces.ed.gov/pubsearch.

Gomes, S. S., Oakes, J., Ormseth, T., Bell, R., & Camp, P. (1990). Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science. RAND. Washington, DC: National Science Foundation.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31*(4), 392–415.

Gottfried, M. A. (2010). Evaluating the relationship between student attendance and achievement in urban elementary and middle schools: An instrumental variables approach. *American Educational Research Journal, 47*(2), 434–465.

Gottfried, M. A. (2014). Chronic absenteeism and its effects on students' academic and socioemotional outcomes. *Journal of Education for Students Placed at Risk, 19*(2), 53–75.

Gottfried, M. A. (2015). Chronic absenteeism in the classroom context: Effects on achievement. *Urban Education*, 1–32.

Gottfried, M. A., & Kirksey, J. J. (2017). "When" students miss school: The role of timing of absenteeism on students' test performance. *Educational Researcher, 46*(3), 119–130.

Gottfried, M., Stiefel, L., Schwartz, A. E., & Hopkins, B. (2017). Showing up : Disparities in chronic absenteeism between students with and without disabilities.

Grissom, J. A. (2011). Can good principals keep teachers in disadvantaged schools? Linking principal effectiveness to teacher satisfaction and turnover in hard-to-staff environments. *Teachers College Record*, *113*(11), 2552-2585.

Grissom, J. A., & Loeb, S. (2011). Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Educational Research Journal* (Vol. 48).

Grissom, J. A., & Redding, C. (2016). Discretion and disproportionality. *AERA Open*, *2*(1), 233285841562217.

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher*, *42*(8), 433–444.

Grissom, J.A., Redding, C. (2016). Discretion and disproportionality: Explaining the underrepresentation of high-achieving students of color in gifted programs.

Guryan, J. (2004). Desegregation and black dropout rates. *American Economic Review*, *94*(4), 919-943.

Guthrie, J.W. (1993). Do America's schools need a 'Dow Jones Index'? *The Phi Delta Kappan*, 74(7), 523-528.

Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness : 1980 - 1995. *School Effectiveness and School Improvement*, *9*(2l), 151–191.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education*, *85*(2), 130-146.

Hanselman, P., & Fiel, J. E. (2017). School opportunity hoarding? Racial segregation and access to high growth schools. *Social Forces*, *95*(3), 1077-1104.

Hanushek, E. A., & Rivkin, S. G. (2008). Do disadvantaged urban schools lose their best Teachers? Brief 7. *National Center for Analysis of Longitudinal Data in Education Research*.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, *4*(1), 131–157.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2009). New evidence about Brown v. Board of

Education: The complex effects of school racial composition on achievement. *Journal of Labor Economics*, *27*(3), 349-383.

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

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, *91*(2), 145-149.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, *24*(3), 411-482.

Heppen, J. B., Walters, K., Clements, M., Faria, A.-M., Tobey, C., Sorensen, N., & Culp, K. (2011). *Access to algebra I: The effects of online mathematics for grade 8 students*. US Department of Education. Retrieved from papers3://publication/uuid/48CED158-1852-495F-BDE9-EC86D889053A

Heyns, B. (1987). Schooling and cognitive development: Is there a season for learning? *Child Development*, 1151-1160.

Holland, A., & Andre, T. (2018). Participation in extracurricular activities in secondary school : What is known, what needs to be known? *Review of Educational Research*, *57*(4), 437–466.

Holme, J. J., Jabbar, H., Germain, E., & Dinning, J. (2017). Rethinking teacher turnover: Longitudinal measures of instability in schools. *Educational Researcher*, 0013189X1773581.

Holt, Stephen B., Gershenson, Seth. (2015). The impact of teacher demographic representation on student attendance and suspensions. Unpublished working paper.

Horng, E., Kalogrides, D., & Loeb, S. (2009). Principal preferences and the unequal distribution of principals across schools. *Center for Analysis of Longitudinal Data in Education Research Working Paper*, *36*(December).

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation (No. w7867). National Bureau of Economic Research.

Innes, J. E. (1989). Disappointments and legacies of social indicators. *Journal of Public Policy*, *9*(4), 429-432.

Isenberg, E., Max, J., Gleason, P., Potamites, L., Santillano, R., Hock, H., & Hansen, M. (2013). Access to effective teaching for disadvantaged students. *Institute of Education Sciences*, (November), 133.

Jackson, C. K. (2009). Student demographics, teacher Sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics*, *27*(2), 213–256.

Jackson, C. K. (2014). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. *NBER Working Paper*, 1–52.

Jacob, B. A., & Lovett, K. (2017). Chronic absenteeism: An old problem in search of new answers. Washington, D.C.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education*, *83*(2), 135–159.

Johnson, M. K., Crosnoe, R., & Elder Jr, G. H. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociology of Education*, 318-340.

Johnson, R. C. (2011). *Long-run impacts of school desegregation & school quality on adult attainments*. NBER Working Paper 16664. http://www.nber.org/papers/w16664.pdf

Johnson, R. C. (2012). The grandchildren of Brown: The long legacy of school desegregation. *Unpublished manuscript, Goldman School of Public Policy, University of California, Berkeley*.

Justman, M., & Méndez, S. J. (2018). Gendered choices of STEM subjects for matriculation are not driven by prior differences in mathematical achievement. *Economics of Education Review*.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research*, (14607), 51.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Research Paper. MET Project. Bill & Melinda Gates Foundation*, 1–48. Retrieved from www.gatesfoundation.org

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*(1), 417–442.

Koch, A. K. (2017). It's about the gateway courses: Defining and contextualizing the issue. *New Directions for Higher Education*, *2017*(180), 11–17.

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*(3), 560–572. https://doi.org/10.1016/j.jue.2008.06.004

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

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

Kraft, M. A. (n.d.). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*.

Kraft, M. A., & Grace, S. (2016). Teaching for tomorrow's economy? Teacher effects on complex cognitive skills and social-emotional competencies. *(Working Paper).* Providence, RI: Brown University.

Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476–500.

Ladd, H. F. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis*, *33*(2), 235–261.

Ladd, H. F., & Sorensen, L. C. (2012). School based accountability and the distribution of teacher quality among grades in elementary schools. Working Paper 75. *National Center for Analysis of Longitudinal Data in Education Research*.

Lam, S. F., Jimerson, S., Wong, B. P., Kikas, E., Shin, H., Veiga, F. H., ... & Stanculescu, E. (2014). Understanding and measuring student engagement in school: The results of an international study from 12 countries. *School Psychology Quarterly*, *29*(2), 213.

Land, K. C. (1975). Theories, models and indicators of social change. *International Social Science Journal*.

Lareau, A. (2003). *Unequal childhoods: Race, class, and family life*. Berkeley, CA: University of California Press.

Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin kindergarten.* Washington, DC: Economic

Policy Institute.

Legewie, J. & DiPrete, T.A. (2012). School context and the gender gap in educational achievement. *American Sociological Review*, 77(3), 463-485.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, *3*(1), 101-28.

Lindsay, C. A., & Hart, C. M. (2017). Teacher race and school discipline. *Education Next*, *17*(1).

Lipscomb, S. (2007). Secondary school extracurricular involvement and academic achievement: a fixed effects approach. *Economics of Education Review*, *26*(4), 463–472.

Little, J. W. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal*, *19*(3), 325–340.

Long, M. C., Conger, D., & Iatarola, P. (2012). Effects of high school course-taking on secondary and postsecondary success. *American Educational Research Journal*, *49*(2), 285–322.

Loveless, T. (2008). The misplaced math student. *The 2008 Brown Center Report on American Education*: *How well are American students learning?*

Loveless, T. (2013). *The 2013 Brown Center Report on American Education*: *How well are American students learning?* Washington, DC: The Brookings Institution. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2001/9/education/09education.pdf

Lutz, B. (2011). The end of court-ordered desegregation. *American Economic Journal: Economic Policy*, *3*(2), 130-68.

Martin, A. J., Mansour, M., Anderson, M., Gibson, R., Liem, G. A. D., & Sudmalis, D. (2013). The role of arts participation in students' academic and nonacademic outcomes: A longitudinal study of school, home, and community factors. *Journal of Educational Psychology*, *105*(3), 709–727. https://doi.org/10.1037/a0032795

McCluskey, C.P., Bynum, T.S. & Patchin, J.W. (2004). Reducing chronic absenteeism: An assessment of an early truancy initiative. *Crime & Delinquency, 50*(2), 214-234.

McEachin, A., Domina, T., & Penner, A. M. (2017). *Understanding the effects of middle school algebra: A regression discontinuity approach*. RAND Working Paper 1209 (Vol. October). Retrieved from www.asdk12.org/forms/uploads/MSprogram.pdf

McKellar, S. E., Marchand, A. D., Diemer, M. A., Malanchuk, O., & Eccles, J. S. (2018). Threats and Supports to Female Students' Math Beliefs and Achievement. *Journal of Research on Adolescence*. Published online first: https://onlinelibrary.wiley.com/doi/abs/10.1111/jora.12384

Meier, K. J. (1993). Latinos and representative bureaucracy testing the Thompson and Henderson hypotheses. *Journal of Public Administration Research and Theory, 3*(4), 393-414.

Mickelson, R. A., Bottia, M. C., & Lambert, R. (2013). Effects of school racial composition on K–12 mathematics outcomes: A metaregression analysis. *Review of Educational Research, 83*(1), 121-158.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. *MET Project Research Paper*, 1–51. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2551&filename=MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Miller, A. (2013). Principal turnover and student achievement. *Economics of Education Review, 36*, 60–72.

Morrissey, T. W., Hutchison, L., & Winsler, A. (2014). Family income, school attendance, and academic achievement in elementary school. *Developmental Psychology, 50*(3), 741–753.

Mueser, P. R. (1979). The effects of noncognitive traits. *Who gets ahead*, 122-58.

Murnane, R. J. (1987). Interpreting the evidence on school effectiveness. *Teachers College Record, 83(1),* 19-35.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills. Principles for educating children to thrive in a changing economy*. New York, NY: Free Press.

Murnane, R. J., Willett, J. B., Bub, K. L., McCartney, K., Hanushek, E., & Maynard, R. (2006). Understanding trends in the black-white achievement gaps during the first years of school. *Brookings-Wharton papers on urban affairs*, 97-135.

National Research Council. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century.* Commit- tee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino and M.L. Hilton, Editors. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nord, C., Roey, S., Perkins, R., Lyons, M., Lemanski, N., Brown, J., and Schuknecht, J. (2011). *Nation's report card: America's high school graduates* (NCES 2011462). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality.* New Haven, CT: Yale University Press.

Oakes, J. (1989). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis, 11*(2), 181-199.

Olson Jr, M. (1969). Social indicators and social accounts. *Socio-Economic Planning Sciences*, *2*(2-4), 335-346.

Orfield, G., Kucsera, J., & Siegel-Hawley, G. (2012). "E Pluribus"... Separation: Deepening double segregation for more students. *Civil Rights Project/Proyecto Derechos Civiles*.

Page, L. C., Murnane, R. J., & Willett, J. B. (2008). *Trends in the Black-White achievement gap: Clarifying the meaning of within-and between-school achievement gaps* (No. w14213). National Bureau of Economic Research.

Palardy, G. J. (2013). High school socioeconomic segregation and student attainment. *American Educational Research Journal*, *50*(4), 714-754.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics, 130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Patillo-McCoy, M. (2000). Negotiating adolescence in a black middle-class neighborhood. *Coping with poverty: The social contexts of neighborhoods, work, and family in the African-American community*, 77-103.

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher*, *42*(5), 259-265.

Penner, A. M., Domina, T., Penner, E. K., & Conley, A. M. (2015). Curricular policy as a collective effects problem: A distributional approach. *Social Science Research*, *52*, 627–641. https://doi.org/10.1016/j.ssresearch.2015.03.008

Pitts, D. W. (2007). Representative bureaucracy, ethnicity, and public schools: Examining the

link between representation and performance. *Administration & Society*, *39*(4), 497-526.

Pounder, D. G., Ogawa, R. T., & Adams, A. E. (1995). Leadership as an organization-wide phenomena: Its impact on school performance. *Educational Administration Quarterly*, *31*(4), 564–588. https://doi.org/0803973233

Rahman, T., Fox, M.A., Ikoma, S., and Gray, L. (2017). *Certification status and experience of U.S. public school teachers: Variations across student subgroups* (NCES 2017-056). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Raudenbush, S. W., & Eschmann, R. D. (2015). Does schooling increase or reduce social inequality? *Annual Review of Sociology*, *41*(1), 443–470. https://doi.org/10.1146/annurev-soc-071913-043406

Ready, D. D. (2010). Socioeconomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure. *Sociology of Education*, *83*(4), 271–286.

Reardon, S. F., & Owens, A. (2014). 60 years after Brown: Trends and consequences of school segregation. *Annual Review of Sociology*, *40*, 199-218.

Reardon, S. F. (2016). School segregation and racial academic achievement gaps. *The Russell Sage Foundation Journal of the Social Sciences*, *2*(5), 34–57.

Reber, S. J. (2010). School desegregation and educational attainment for blacks. *Journal of Human Resources*, *45*(4), 893-914.

Rees, D. I., & Sabia, J. J. (2010). Sports participation and academic performance: Evidence from the National Longitudinal Study of Adolescent Health. *Economics of Education Review*, *29*(5), 751–759.

Richtman, K.S. (2007). The truancy intervention program of the Ramsey County Attorney's office: A collaborative approach to school success. *Family Court Review*, 45(3), 421-437.

Riegle-Crumb, C., King, B., Grodsky, E., & Muller, C. (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. *American Educational Research Journal*, *49*(6), 1048-1073.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4–36.

Rothstein, J., & Mathis, W.J. (2013). Review of "Have we identified effective teachers?" and "A composite estimator of effective teaching: Culminating findings from the measures of effective teaching project". *National Education Policy Center,* 13.

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record, 107*(9), 1999.

Ruzek, E. A., Domina, T., Conley, A. M. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *Journal of Early Adolescence* (Vol. 35).

Saatcioglu, A. (2010). Disentangling school-and student-level effects of desegregation and resegregation on the dropout problem in urban high schools: Evidence from the Cleveland Municipal School District, 1977-1998. *Teachers College Record, 112*(5), 1391-1442.

Schwartz, H. (2010). *Housing policy is school policy: Economically integrative housing promotes academic success in Montgomery County, Maryland*. Century Foundation.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven, CT: Yale University Press.

Seashore-Louis, K., Leithwood, K., Wahlstrom, K.L., & Anderson, S.E. (2010). Learning from leadership: Investigating the links to improved student learning. *ERS Informed Educator, 2012*(10/7/2012), 1–11.

Sharkey, P. (2010). The acute effect of local homicides on children's cognitive performance. *Proceedings of the National Academy of Sciences, 107*(26), 11733-11738.

Sheldon, S.B. (2007). Improving student attendance with school, family, and community partnerships. *The Journal of Education Research, 100*(5), 267-275.

Sheldon, S.B. & Epstein, J.L. (2004). Getting students to school: Using family and community involvement to reduce chronic absenteeism. *School Community Journal, 14*(2), 39-56.

Simzar, R., Domina, T., & Tran, C. (2016). Eighth-grade algebra course placement and student motivation for mathematics. *AERA Open, 2*(1), 233285841562522.

Sinclair, M. F., Christenson, S. L., Lehr, C. A., & Anderson, A. R. (2003). Facilitating student engagement: Lessons learned from Check & Connect longitudinal studies. *The California School Psychologist*, *8*(1), 29-41.

Smerillo, N. E., Reynolds, A. J., Temple, J. A., & Ou, S. R. (2018). Chronic absence, eighth-grade achievement, and high school attainment in the Chicago Longitudinal Study. *Journal of School Psychology*, *67*(October 2016), 163–178.

Smith, M. S. (1988). Educational indicators. *The Phi Delta Kappan*, 69(7), 487-491.

Smith, J., Hurwitz, M., & Avery, C. (2017). Giving college credit where it is due: Advanced Placement exam scores and college outcomes. *Journal of Labor Economics*, *35*(1), 67–147.

Stearns, E., & Glennie, E. J. (2010). Opportunities to participate: Extracurricular activities' distribution across and academic correlates in high schools. *Social Science Research*, *39*(2), 296–309.

Stempel, H., Cox-Martin, M., Bronsert, M., Dickinson, L. M., & Allison, M. A. (2017). Chronic school absenteeism and the role of adverse childhood experiences. *Academic Pediatrics*, *17*(8), 837–843.

Subramanian, S. V., & Kennedy, M. H. (2009). Perception of neighborhood safety and reported childhood lifetime asthma in the United States (US): A study based on a national survey. *PloS one*, *4*(6), e6091.

Supovitz, J., Sirinides, P., & May, H. (2010). How principals and peers influence teaching and learning. *Educational Administration Quarterly*, *46*(1), 31–56.

Taie, S. and Goldring, R. (2017). Characteristics of public elementary and secondary school principals in the United States: Results from the 2015–16 National Teacher and Principal Survey. First Look. NCES 2017. *National Center for Education Statistics*.

Thompson, K. D. (2017). What blocks the gate? Exploring current and former English learners' math course-taking in secondary school. *American Educational Research Journal*, *54*(4), 757–798.

Valentine, J. C., Konstantopoulos, S., & Goldrick-Rab, S. (2017). What happens to students placed into developmental education? A meta-analysis of regression discontinuity studies. *Review of Educational Research*, *87*(4), 806–833.

Van Knippenberg, D., Dahlander, L., Haas, M. R., & George, G. (2015). Information, attention, and decision making. *Academy of Management Journal*, *58*(3), 649-657.

Vigdor, J.L. (2011). School desegregation and the black-white test score gap. In G.J. Duncan & R.J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances,* (pp. 443-464). New York, NY: Russell Sage Foundation.

Wang, M.T. & Eccles, J.S. (2013). School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction,* 28, 12-23.

Weiner, D. A., Lutz, B. F., & Ludwig, J. (2009). *The effects of school desegregation on crime.* NBER Working Paper 15380. Cambridge: NBER.

Wells, A. S., & Crain, R. L. (1994). Perpetuation theory and the long-term effects of school desegregation. *Review of Educational Research*, *64*(4), 531-555.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*, 61–78.