Policy Notes

EQUITY IN EDUCATION SERIES

Indicators of Teaching Quality:

Appraising the Case of Early Childhood Education

Gary Sykes
Courtney Bell
Bhavya Shukla

THE ETS CENTER FOR RESEARCH ON HUMAN CAPITAL AND EDUCATION

Table of Contents

This report was written by:

Gary Sykes
Courtney Bell
Bhavya Shukla

The analysis reported in this policy note was made possible in part by a grant from the Spencer Foundation (#201700160). The views expressed are those of the authors and do not necessarily reflect the views of the Spencer Foundation.

March 2020

ETS Center for Research on Human Capital and Education

Research and Development
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001

Suggested citation: Gary Sykes, Courtney Bell, and Bhavya Shukla, *Indicators of Teaching Quality: Appraising the Case of Early Childhood Education* (Princeton, NJ: Educational Testing Service, 2020).

Introduction

> *We can't fix what we can't see.*
>
> *— a NASA® official regarding fine-grained satellite images of Earth*

Americans, it might be said, accept a good deal on faith, but they also place a lot of faith in numbers. Measures are used in almost every field of endeavor today, including commerce, politics, science, and popular culture (think fantasy sports leagues). In fields such as medicine, measurement has proven crucial to progress in diagnosis and treatment. Hospitals and organizations of all kinds use measurement to support decision making. Leaders in business and government routinely rely on a range of indicators to guide problem analysis and strategy. Browse among offerings from any of the management gurus at your nearby airport bookstore and you will find a lot of attention devoted to measuring what matters.

The education field is no exception to the growing importance of measurement, abetted by dramatic advances in computing and data analysis. One prominent measure, the National Assessment of Educational Progress (NAEP), now celebrates 50 years of reporting on the academic achievement of America's students, and many other kinds of educational measures have been developed and used over the years. Through programs such as NAEP and other recurring data collections, the federal government tracks indicators on outputs as far as what students are learning in school, and on trends in inputs to schooling such as per-pupil expenditures, school staffing, and educator credentials.

However, in K–12 education, there is an anomaly to this story of progress guided by advances in the science of measurement. There is a hole right between the inputs and outputs of education. We have no valid, widely used measures of teaching itself, of the quality of instruction across the public education system. This is a crucial omission because without an understanding of the instructional process, we have no firm basis for interpreting trends in student learning or, more importantly, for guiding efforts at improvement—we can't fix what we can't see. Measures of inputs are important for many purposes, but they are weak proxies for what teachers do with students in the classroom on a daily basis, and results from scores of studies of student achievement clearly indicate that classroom-level inputs are what matters for schooling's contribution to outcomes. As for these outcomes, to be clear, even the best measures do not tell us anything about how to achieve the best results.

Ideally, we would have reliable information that allows for tracking trends over time on what goes on in the classroom. From studies culled over many years, we have indications about instruction that mirror the inequality seen in outcomes, pointing to troubling tendencies. Inequality shows up on every measure in use, suggesting strongly that instructional quality is inequitably provided. Variation across multiple measures, as many studies have shown, is significant not only between schools but also within schools, typically to the disadvantage of minoritized students.[1] Notwithstanding the absence of systematic trend data, the portrait of instruction rising out of many individual studies paints a longstanding picture of dull mediocrity, low-level academic curricula, and bored, disengaged students.[2] There are, of course, many exceptions to these broad generalizations, and the American school system has

made a significant contribution historically to the welfare of the nation. However, the educational past is no longer prologue to a promising future. Progress will rely on qualitative improvements that can be scaled up and implemented across many states, communities, districts, schools, and classrooms. Such progress undoubtedly will be hindered if we cannot develop better measures of the quality of teaching, grounded in the daily interactions among students and teachers. Since we can't fix what we can't see, we must use these measures to improve the quality of instruction in reliable ways—in short, to fix what we *can* see.

This is the second in a series of policy notes and related publications exploring prospects for developing indicators of teaching quality.[3] The first examined federal databases that include information about teachers, demonstrating that almost none of the measures in these datasets provides indication of teaching quality. This policy note takes the next step by concentrating on measures of instruction associated with valued outcomes for students, using a tripartite set of indicators—structural, process, and outcome—to guide policy and decision making. After an overview about indicators in general, we look at indicators specifically as they apply to the K–12 education sector, discussing how there is no shortage of structural (inputs) and outcome (assessments of learning) indicators before delving into the gap in process indicators (measures of quality of instruction). We then explore one example of how such indicators of instruction have been developed and used in early childhood education (ECE) in regard to the federal Head Start program.

On Indicators

An indicator is a statistic or combination of statistics (e.g., composing an index) designed to gauge progress toward some important end or outcome. Unlike accountability measures, indicators have no stakes attached to them and are used primarily to track trends in matters regarded to have intrinsic or instrumental value, particularly in relation to important outcomes. Indicators are familiar in many sectors, including, for example, the unemployment rate or the Dow Jones® average of industrial stocks.

Three kinds of indicators have been developed in many fields today. One kind of indicator—referred to as "structural"—tracks inputs, such as the licensure and certification status of healthcare professionals. When seeking medical help, for example, patients might check to see if the physician is board certified in the relevant medical specialization. In education, a widely used structural indicator is class size. It is valued by parents, and studies have shown it is related to learning outcomes.

Another kind of indicator measures outputs or outcomes. Hospitals, for example, track mortality and morbidity rates, while achievement test scores are now in wide use in education. Increasingly, educational outcomes include not only academic achievement, but also social and emotional learning.

Finally—and this is where our analysis centers—indicators might measure the quality of services provided by professionals. These "process indicators" typically measure features of interactions among caregivers and their patients or clients. For example, process measures for quality in surgical care range from relatively modest indicators such as use of preanesthesia checklists to more robust indicators such as use of perioperative beta blockers in patients at risk for cardiac events and appropriate use of antibiotics for wound prevention.[4] For educational process indicators, instruction in the classroom is a focal point. Process indicators have proven difficult to define and measure, but, in general, they enjoy a closer relationship to valued outcomes than do structural indicators.

Indicator development has followed a particular trajectory. During the founding of institutions, structural indicators are developed first in order to track common, easily measured inputs. As hospitals, schools, and universities emerged, the focal point was on institution building. Such matters as per-pupil expenditures, supply of well-provisioned schools, and qualifications of key personnel have accompanied the creation of school systems. Somewhat later in the process, measures of outcomes also took shape to determine if schools were fulfilling their purposes. NAEP, for example, was first administered some 120 years after the early formation of America's schools. Typically, modest associations have been shown between structural and outcome indicators, with improvement oriented around the structural indicators. Gradually, for example, educational qualifications for teachers, nurses, and other human-service workers were increased, with such increases justified on the basis of their relation to service delivery and outcomes.

As institutional sectors emerged and matured, concerns arose about the quality of services provided. Service quality was considered important because of its intrinsic value, its relation to outcomes, and its use in processes of continuous improvement of services. This final point is particularly important. Structural indicators provide little clue on how to improve service delivery, while outcomes merely provide for assessment of the status of those being measured and signal where improvement might be needed. The challenge has been to develop process indicators that avoid undesirable responses, are reliably related to valued outcomes, are cost efficient and minimally burdensome to respondents, and remain stable over time to permit trend analysis.[5]

In mature fields such as medicine, the three types of indicators work together to ensure service delivery and quality of care.[6] In the education field, development has been slower and less sure. The case of Head Start, which we turn to next, offers an example, illustrating the historical trajectory. Ultimately, we are interested in the prospects for indicators in the K–12 sector, while the Head Start case provides a potentially suggestive lead.

A Case in Point: Indicators in Early Childhood Education

Today, 65 percent of 4-year-old children are enrolled in some type of center-based care, including child care, Head Start, or prekindergarten.[7] The story of how indicators of quality entered the ECE policy stream illustrates both the perils and promises of grounding policy and practice in research. We offer an abbreviated account to draw some implications for indicator development in the K–12 sector.

A now-famous project and the accompanying studies helped put ECE on the policy map. The High Scope/Perry Preschool Project, launched in Ypsilanti, Michigan, in 1962, has been extensively studied. Children who participated have been tracked over time on a variety of academic and life outcomes. Findings have been qualified in various ways over the years, but overall, the program achieved notable success for the children who were enrolled.[8] Both short- and long-term effects provided persuasive evidence that children, particularly those from low-income households, benefited from education in the early years of life.[9] Simply extending ECE across the states to cover more children appeared to be the first priority for the state-funded pre-K programs. What soon became apparent was the great variability in the programs that were offered. How to introduce some form of quality control—some means of stimulating improvement—became a critical issue.

The initial approach involved development of a protocol for direct observation in classrooms, which has undergone several revisions. The original protocol, known as the Early Childhood Environmental Rating Scale (ECERS), concentrated primarily on such structural inputs as books, sand and water tables, and features promoting the health and safety of children. Then, a revised version, the ECERS-R, expanded attention to other matters such as "space and furnishing," "parents and staff," "activities," and "interactions." Evidence indicates that ECERS-R scores have improved gradually over the years, while at the same time, links between ECERS-R scores and child outcomes have weakened.[10]

Efforts to improve the scale have continued. A group of educational researchers in the United Kingdom developed four additional curricular subscales that measure literacy, numeracy, science, and diversity known as ECERS-E.[11] For the most part, only the ECERS-E extension includes attention to process measures. It has not yet been carefully validated and is not in widespread use. However, another version, ECERS-3, has come into use, featuring greater attention to process indicators.

Also widely used in settings under state jurisdiction is the quality rating and improvement system (QRIS), which aggregates several separate indicators of quality into a composite rating designed to provide parents and policymakers with a guide to quality programs. QRIS indicators include such items as staff qualifications, health-and-safety features, and staff-child ratios. QRIS expanded very rapidly in response to Race to the Top Early Learning Challenge Grants during the Obama administration. All 50 states are now at some stage of implementing a QRIS initiative. However, evidence supporting a relation between QRIS and child outcomes is not encouraging, leading to a judgment that the "lack of predictive validity is a serious threat to the utility of QRIS."[12] Moreover, one critique of QRIS systems is their lack of clarity regarding the actual ways in which a program may achieve a certain rating. The approaches to aggregating indicators (e.g., using cut points on continuous measures, assigning weights to various components, how indicators roll up to an aggregate) differ across QRIS systems, making it difficult to discern how to interpret a given ranking (e.g., a "three star" program) or to use QRIS information to drive investments in improvement.

Bringing CLASS to Class

In parallel to the work described above, a robust research and theoretical basis gradually emerged for understanding that factors associated with the qualities of interaction between teachers and children in early childhood settings were key contributors to children's social, emotional, and academic outcomes.[13] Here, the story about indicators benefits from a serendipitous confluence of events. As Congress was contemplating the reauthorization of Public Law 110-134, the Improving Head Start for School Readiness Act in 2007, pressures had begun to mount for greater accountability. Legislators wanted to see that the program was achieving results. An initial effort to test children directly proved unpopular and impractical, sparking a search for alternatives. One option was direct observation of classrooms, and the instrument in use at the time was ECERS-R. But the relation of ECERS-R scores to outcomes was weak. Furthermore, most Head Start programs ended up with ratings of above average.

At the same time, another observation system, the Classroom Assessment Scoring System (CLASS)®[14] had been developed and tested by a group of scholars primarily at the University of Virginia. CLASS orients attention to three broad domains of classroom interactions hypothesized to be important in promoting student learning and social development. Table 1 supplies details on this conceptualization of the factors contributing to instructional quality.

Table 1: Teaching through Interactions Framework: Description of Domains and Dimensions

DOMAIN	DIMENSION	DESCRIPTION
EMOTIONAL SUPPORT	Positive Climate	Reflects the overall emotional tone of the classroom and the connection between teachers and students
	Negative Climate	Reflects overall level of expressed negativity in the classroom between teachers and students (e.g., anger, aggression, irritability)
	Teacher Sensitivity	Encompasses teachers' responsivity to students' needs and awareness of students' level of academic and emotional functioning
	Regard for Student Perspectives	The degree to which the teacher's interactions with students and classroom activities places an emphasis on students' interests, motivations, and points of view, rather than being very teacher-driven
	Overcontrol	Assesses the extent to which the classroom is rigidly structured or regimented at the expense of children's interests and/or needs
CLASSROOM ORGANIZATION	Behavior Management	Encompasses teachers' ability to use effective methods to prevent and redirect misbehavior by presenting clear behavioral expectations and minimizing time spent on behavioral issues
	Productivity	Considers how well teachers manage instructional time and routines so that students have the maximum number of opportunities to learn
	Instructional Learning Formats	The degree to which teachers maximize students' engagement and ability to learn by providing interesting activities, instruction, centers, and materials
	Classroom Chaos	The degree to which teachers ineffectively manage children in the classroom so that disruption and chaos predominate
INSTRUCTIONAL SUPPORT	Concept Development	The degree to which instructional discussions and activities promote students' higher-order thinking skills versus focus on rote and fact-based learning
	Quality of Feedback	Considers teachers' provision of feedback focused on expanding learning and understanding (formative evaluation), not correctness or the end product (summative evaluation)
	Language Modeling	The quality and amount of teachers' use of language-stimulation and language-facilitation techniques during individual, small-group, and large-group interactions with children
	Richness of Instructional Methods	The extent to which teachers use a variety of strategies to promote children's thinking and understanding of material at a deeper and more complex level

In CLASS, specific, observable behavioral descriptions are anchored at three points along a seven-point rating scale, with some substantive changes across grades so that the measured constructs are developmentally appropriate.

In 2008, a study was released based on the use of both CLASS and ECERS-R in an 11-state sample.[15] Results demonstrated that CLASS was associated with academic and language skills together with teacher-supported social skills, while ECERS-R had a much weaker relationship, predicting only one of the set of outcomes.[16]

The combination of pressures for accountability and the positive results from this and related studies led to the inclusion of CLASS as one of the accountability instruments in the reauthorization legislation. Among other aspects of program administration and provision, Head Start grantees now also would be accountable for their results on CLASS in the course of their mandated three-year cycle of reviews. If results were poor, a program would need to make improvements and reapply for funding.

Implementation challenges were formidable in the early going. If observation was to serve as the measurement mode for an indicator system, a host of implementation issues—and their costs—needed to be managed. Raters had to be trained, and ratings had to be collected, stored, and analyzed. In this case, serendipitously, the Virginia team had considerable experience in this area from large-scale research studies where it had conducted classroom observations using a large number of observers following a standardized protocol (using CLASS) in a wide range of settings dispersed across the country. The developers used a "train the trainer" model to prepare raters, utilizing the wealth of resources already created by the Virginia team, including training protocols, video exemplars, and a cadre of experienced raters.

Eventually, Head Start contracted with Danya International, a firm that provides a range of services associated with data management, program monitoring, training, and technical assistance, to manage the training and certification of raters along with other responsibilities. Danya was charged with the triennial reviews of nearly 1,700 Head Start grantees across the country, eventually assuming responsibility for training and certifying raters to conduct the classroom observations with the CLASS instrument. Danya provided a range of services, funded by the Office of Head Start (OHS), that included rater training, data collection storage and retrieval, report preparation, and reporting results to OHS and the grantees. In short, it was possible to scale a standardized classroom observation protocol across the entire country.

Today, CLASS is used extensively in the Head Start programs as an accountability measure and a focal point for improvement and professional development, including in 23 state QRIS's.[17] This use case demonstrates the feasibility of direct observations of interactions in the classroom in a standardized manner that ensures consistency and reliability of scores. At present, some 35,000 observers have been trained in one or more versions of CLASS, totaling more than 70,000 certifications.[18] Critical to this enterprise has been the effort to standardize observation through three components: (1) a training protocol; (2) parameters for conducting observations; and (3) directions for scoring.[19] The developers created comprehensive materials to accompany the training protocol. They included a detailed training manual, videos and transcripts with gold standard scores for scoring practice, procedures for reliability checks, and guidelines for completing the training before using the observation tool, along with resources (video, online support) that recertify observers on a periodic basis and guard against observer drift and bias. Several of the important parameters included length of observations, start and stop times, directions for time of day and specific

activities to observe, whether to announce observations beforehand, and related issues. A number of these parameters have been the focus of explicitly empirical analysis, using extant data and specifically designed studies.

In addition to such standardization guidelines, both reliability and validity of scores were important to establish. In particular, evidence had to track stability over time in ratings and consistency across observers. Concerns about validity centered particularly on whether observation scores predicted outcomes of interest, so studies to evaluate the relations between CLASS and outcomes were necessary. Evidence across studies has estimated a modest but consistent relationship between CLASS ratings and child outcomes both in U.S. and international samples.[20]

The developers of CLASS attended to these and related issues over an extended period, demonstrating the feasibility of observing quality at scale. They also showed the importance of a parallel research program that evaluated the multiple questions that arise from implementation at scale. Because CLASS had been adopted in Head Start as an accountability measure, programs had strong incentives to attend to their periodic CLASS scores, which determined eligibility to reapply for funding.[21] Also, the Head Start program supplied funding and support for the infrastructure needed to support the observation system.

Still, does the record show steady improvement in programs and their results over time as a consequence of attending to the indicators of quality represented in CLASS? The critical issue concerned how CLASS was integrated into, and aligned with, the improvement process. At the outset, the architects of CLASS recognized the need for a strategy to improve classroom quality and outcomes, so they devised and evaluated a professional-development program for teachers that was aligned with the assessment system. Teachers could work on elements of their instruction where CLASS ratings indicated a need for improvement. Through guided analysis of video clips of their own interactions together with access to exemplars, teachers could work systematically on aspects of their practice that the ratings revealed as areas of weakness. A series of experimental studies of the professional-development program indicated the approach was effective for improving the quality of teacher-child interactions and children's school readiness.[22]

Soon after implementing CLASS as part of the triennial accountability review, Head Start made technical assistance and other resources available to programs to improve quality. However, it did not specify or restrict attention to the professional development that had a proven track record of success, and furthermore, these investments were not targeted to specific forms of quality. In many instances, the professional development was allocated to local priorities that had no particular research backing, unlike the program associated with CLASS that studies have demonstrated is effective in improving learning and development outcomes. Moreover, although some professional-development resources and approaches flowed to programs as a function of focused training and technical assistance goals, use of other resources was left to the discretion of local programs.

As a result, resources for improvement were allocated around a wide range of local priorities and offerings, weakening the link—at both the program and Head Start system levels—between the observed indicators of program quality and the professional-development strategies and investments presumably being deployed toward improvement. Reflecting on this history, Robert Pianta, a chief architect of CLASS, commented that all of the critical elements involved in systematic improvement were never "bolted together," so the ultimate promise of the CLASS may not have been fully realized, at least to date.[23] However,

two other use cases of CLASS as a core indicator of quality within an improvement-focused framework demonstrate the benefits of tightly linking quality indicators with improvement strategies and approaches.

In the first use case, the Dallas Independent School District adopted CLASS for use in pre-K, kindergarten, and first-grade classrooms across the district and employed professional-development models aligned to CLASS. A recent study evaluating this approach followed children from pre-K through first grade to determine whether they were "on track" in terms of academic outcomes.[24] Children in classrooms with high CLASS ratings were almost three times as likely to be on track as those who were not in pre-K programs and were exposed to ineffective teaching in kindergarten and first grade, exposing a very significant "opportunity gap." The district then mobilized professional development for teachers, including CLASS-aligned coursework, coaching, and use of video exemplars in cycles of feedback and trial. While in 2015, only 40 percent of pre-K teachers were providing effective instructional support to children (as measured on CLASS), by fall 2017, that number was up to 60 percent. When the pieces were bolted together, results followed.

The second use case involves the state of Louisiana and bears observation. There, state officials made the decision to adopt CLASS as their QRIS indicator, concentrating attention just on the observable dimensions of classroom interactions measured by CLASS. In order to scale up this approach, the state relies on a cadre of local raters who conduct a minimum of two CLASS observations, with half the classrooms moderated by an external, third-party rater whose score prevails in cases of discrepancy.[25] With state support, communities now have over 1,200 reliable CLASS observers. The state is also engaged in providing regular training for the ECE workforce on CLASS, which is becoming a significant source of learning, and in partnering with a team of researchers to ensure rigorous empirical evaluation of policy decisions and implementation parameters. Although evidence on the effects of this new policy approach has yet to be gathered, one result has been documented: Those departing the early-childhood teaching profession have lower CLASS scores than those who stay or are entering.[26] Such a finding suggests that quality is improving gradually over time.

In this use case, CLASS figures in the state accountability system for early learning centers that include Head Start and programs receiving Child Care and Development Block Grant funds, but not public or nonpublic elementary schools that do not need to be licensed unless they serve 3-year-olds. Programs that score poorly on CLASS over a period of years risk losing their funding.

Nationally, significant gaps remain in access to high-quality early childhood education. By one estimate, gaps between public ECE for nonpoor, nonminority children and minoritized children in state-funded pre-K programs ranges from 0.3 to 0.7 standard deviations on a range of observational measures, with the largest gaps observed on CLASS, which is the strongest predictor of child outcomes across student groups.[27] One recent study employing CLASS found that only four percent of children in rural areas of North Carolina and Pennsylvania had access to good enough teaching from kindergarten to third grade, and over 50 percent experienced good enough teaching for only one year or less.[28] However, when students were exposed to quality teaching (as defined by CLASS) for multiple years in a row, notable improvements in achievement were detected. This cumulative effect of quality educational experience is of considerable importance.[29]

It is significant to note that structural classroom characteristics explain little of the gaps in process quality, while characteristics of peers in the classroom explain more than half. There is large between-state variation in access to quality ECE, and within-state residential

segregation is correlated with quality gaps. When Black and Hispanic children are segregated, they tend to be in worse ECE environments than their White peers.[30] Further work on access to high-quality ECE settings and on the ECE workforce is critical to address these gaps.

The CLASS Story in Perspective

This story calls for a reckoning. The common tendency in the education field is to develop measures and interventions for use in particular studies that are rarely deployed afterward. In the case of CLASS, developers created an observational process indicator grounded in theory and research on child development; examined its technical properties through a wide-ranging, extensive set of studies; created a rich set of coordinated supporting resources around the measure; and introduced the whole into the policy system. That arc of development, if not unprecedented, is unusual.

CLASS indicates you can fix what you can see. Unfortunately, the education field has not had a valid and reliable way of seeing the quality of instruction in American classrooms as a basis for effecting improvement. When it comes to quality, without a mechanism for detecting what is signal amid the noise in classrooms, policymakers and educators are hampered in their ability to guide systematic improvements.

The story of CLASS in the ECE sector provides one intimation of the possible. Although policy development benefited from a happy confluence of circumstances, where a persuasive study arrived just in time to influence program reauthorization, the research groundwork had been carefully laid over a period of years. Arguably, the program of development and the policy strategy nearly got to the goal line of delivering systematic improvement, falling short only because the crucial elements ultimately were not tightly bolted together enough to maximize their impact on student learning. However, it can be done, as demonstrated by the emerging evidence in Dallas and Louisiana. Other studies have reached similar conclusions that multiple policy implements must be aligned in order to exert influence on teaching and learning.[31] CLASS clearly illustrates the point.

The case also demonstrates that the costs and logistics associated with measuring teaching quality at scale can be managed. With CLASS, a critical feature involved how the costs of training raters and accounting for their time commitments was folded into the policy model. The obvious analog to the Head Start program in early childhood education is to bolster the Title I program (the federal program for schools with high percentages of children living in poverty) in K–12 education. One can imagine requirements built into Title I regulations that would support data collection efforts for Title I schools, which account for 54 percent of all public schools, with elementary schools accounting for the greatest share. The supporting policy argument would be that we need information the most on instructional quality in the schools serving our least advantaged students.

The CLASS story also makes another point clear. Proxies for direct measures of instruction are a weak basis for policymaking. The measures that count when it comes to assessing the quality of teaching must capture interactions in the classroom among teachers and students around content. Good indicators are good predictors of the outcomes that matter—children's intellectual, social, and emotional development. If developments in the K–12 sector are to follow the lead of ECE, then indicators of instructional process must be developed. Relatedly, bundling together a set of weak proxies, as evidenced in many states' QRIS's, is less effective than concentrating improvement around a parsimonious, clear, predictive set of indicators, as exemplified in CLASS.

In the ECE case, CLASS was not introduced as part of an indicator system. Rather, it was a program accountability measure linked to specific stakes—program approval. Indicators operate differently. They are low-stakes measures, enjoying the accompanying benefits and liabilities. The benefit is that low-stakes measures do not introduce the distortions that typically arise when programs seek any means possible to score well on the measures. The liability is that indicators are not necessarily joined with any pressures to comply and change behavior in the face of a poor showing. Instead, it is up to policymakers to respond when indicators demonstrate poor performance.[32] In the Head Start case, policymakers could insist on accountability because it is a federal program. However, K–12 education is a state function, and federal policymakers must avoid overstepping their authority[33] —an argument for employing measures as indicators rather than for accountability.

Reconsiderations

Still larger forces have been at work undermining access to quality for low-income and minoritized students. Inequalities are stitched into the structures of educational opportunity, whether by design or happenstance. State policies play a significant role in the provision of ECE, and they vary widely. Segregation nearly always lurks, influencing educational processes and outcomes. Intervention on these large matters seems prerequisite to progress.

However, tackling large structural issues is a heavy lift politically in many locales, so the alternative presents itself: Work on improving the current group of teachers. That is the promise offered by the story of CLASS in the ECE sector, but there are those who will argue for macro change as a prerequisite to micro change.

Another skeptical retort to this analysis notes the major differences between the ECE and K–12 educational sectors. K–12 education is considerably larger than ECE, so issues of scale are formidable. Goals for K–12 schooling also may be more expansive and contested, and the age and developmental span of students is greater in K–12 than in ECE. Furthermore, the K–12 sector is more highly differentiated in terms of grade levels, school types, academic subjects, and other features. The markets for tests, textbooks, professional development, and other resources are filled with contending competitors who might resist moves toward standardization. And the regulatory environments differ, with no direct analog to the influence of the local school board in the ECE sector. Additionally, national indicators could run into opposition around the hallowed ideal of local control.

These and other factors certainly combine to increase the challenges involved in developing an indicator system for teaching quality that would provide usable information for K–12 public education. We don't mean to overstate the case. In the face of these challenges, we return ultimately to the fundamental point made by Pianta and Hamre when they write, "It is stunning, given the importance of classroom settings for the transmission of knowledge and skill in our system of education, that little or no population-level data exist pertaining to exposure of children and adolescents to particular classroom practices that are either known to relate to academic success or failure, desired on the basis of certain policies or values, or even hypothetically expected to relate to outcomes."[34] Bringing CLASS to class certainly doesn't settle all of the obvious difficulties, but it opens the door just a bit to imagining the possibilities rendered so crucial by the stakes involved—the education of a nation's most precious resource: its children.

About the Authors

Gary Sykes is a Principal Research Scientist in ETS's Student and Teacher Research Group, where he is responsible for investigating a wide range of measures, instruments, and procedures used to assess and improve teaching quality. His interests focus particularly on issues of equity in teaching; policies related to teachers, teaching, and teacher education; and efforts to enhance the teaching profession.

Courtney Bell, Principal Research Scientist in ETS's Center for Global Assessment, is a former high school science teacher and teacher educator. Her teaching in rural North Carolina convinced her to study the intersections of research, policy, and practice. Her studies use mixed methods to analyze teaching, teacher education, and the validity of teaching quality measures, especially observational measures. Recent studies investigate how administrators learn to use a high-stakes observation protocol, how to understand the validity of observational measures, how measures of teaching compare across countries, and the ways in which observation protocols capture high-quality teaching for students with special needs.

Bhavya Shukla, Senior Research Assistant at ETS, holds a master's degree in Education and Social Policy from New York University. She has primarily worked in the education sector in India on school leadership development programs and a randomized control study on measures for developing intrinsic teacher motivation for secondary school teachers. Her interests lie particularly in equity in K–12 education. Some of her more recent work includes projects on developing an indicator system to evaluate teaching quality in the United States, a special education teacher framework, and global assessment of classroom teaching and learning.

Endnotes

1 Unusual as the term is, we use "minoritized" in referring to children and families of color that do not have access to equitable opportunity even when they constitute the numerical majority within their communities. They remain excluded and relegated to an underprivileged status. The term refers to any group that is devalued in society due to differences in race, ethnicity, immigration status, social class, gender, native language, and other markers of identity. See Bryant Jensen, Sara Grajeda, and Edward Haertel, "Measuring Cultural Dimensions of Classroom Interactions," *Educational Assessment* 23, no. 4 (2018): 250–276, https://doi.org/10.1080/10627197.2018.1515010.

2 A sample of this large literature includes Robert Crosnoe, Fred Morrison, Margaret Burchinal, Robert Pianta, Daniel Keating, Sarah L. Friedman, K. Alison Clarke-Stewart, and The Eunice Kennedy Shriver National Institute of Child Health and Human Development Early Child Care Research Network, "Instruction, Teacher-Student Relations and Math Achievement Trajectories in Elementary School," *Journal of Educational Psychology* 102*,* no. 2 (2010): 407–417, https://dx.doi.org/10.1037%2Fa0017762; Larry Cuban, *How Teachers Taught. Constancy and Change in American Classrooms, 1880–1990,* 2nd Ed. (New York: Teachers College Press, 1993); John I. Goodlad, *A Place Called School: Prospects for the Future* (New York: McGraw-Hill, 1984); James Hiebert, James W. Stigler, Jennifer K. Jacobs, Karen Bogard Givvin, Helen Garnier, Margaret Smith, Hilary Hollingsworth, Alfred Manaster, Diana Wearne, and Ronald Gallimore, "Mathematics Teaching in the United States Today (and Tomorrow): Results from TIMSS 1999 Video Study," *Educational Evaluation and Policy Analysis* 27, no. 2 (2005): 111–132, https://www.jstor.org/stable/3699522; James Hoetker and William P. Ahlbrand Jr., "The Persistence of the Recitation," *American Educational Research Journal* 6, no. 2 (1969): 145–167, https://doi.org/10.3102%2F00028312006002145; Robert Pianta, Carollee Howes, Margaret Burchinal, Donna Bryant, Richard Clifford, Diane Early, and Oscar Barbarin, "Features of Pre-Kindergarten Programs, Classrooms, and Teachers: Do They Predict Observed Classroom Quality and Child-Teacher Interactions?" *Applied Developmental Science* 9, no. 3 (2005): 144–159, https://doi.org/10.1207/s1532480xads0903_2; Ming-Te Wang, Maureen Brinkworth, and Jacquelynne Eccles, "Moderating Effects of Teacher-Student Relationship in Adolescent Trajectories of Emotional And Behavioral Adjustment," *Developmental Psychology* 49, no. 4 (2013): 690–705, https://doi.org/10.1037/a0027916; and Iris R. Weiss, Joan D. Pasley, Sean P. Smith, Eric R. Banilower, and Daniel J. Heck, *Looking Inside the Classroom: A Study of K–12 Mathematics and Science Education in the United States* (Chapel Hill, NC: Horizon Research, 2003).

3 Our first policy note was *Quest for Quality: An Indicator System for Teaching* (Princeton, NJ: Educational Testing Service, 2019), https://www.ets.org/s/research/pdf/quest-for-quality.pdf.

4 John D. Birkmeyer, Justin B. Dimick, and Nancy J. O. Birkmeyer, "Measuring the Quality of Surgical Care: Structure, Process, or Outcomes?" *Journal of the American College of Surgeons* 198, no. 4 (2004): 626–632, https://doi.org/10.1016/j.jamcollsurg.2003.11.017.

5 Richard J. Murnane, "Improving Education Indicators and Economic Indicators: The Same Problems?" *Educational Evaluation and Policy Analysis* 9, no. 2 (1987): 101–116, https://doi.org/10.3102%2F01623737009002101.

6 Avedis Donabedian, "The Quality of Care. How Can It Be Assessed?" *Journal of the American Medical Association* 260, no. 12 (1988): 1743–1748.

7 The sector denoted as "early childhood education" conceals some considerable variation pertinent to this analysis. The care and education of children from birth to the age of 5 includes child care facilities, preschools, state-funded pre-K programs, and the federal Head Start program. These alternatives, including private and publicly funded options, come under the jurisdiction of various state and federal regulatory agencies, each with their own rules and regulations. In the account to follow, we will distinguish how events and developments have affected one or another of these programs.

8 See, for example, Clive R. Belfield, Milagros Nores, Steve Barnett, and Lawrence Schweinhart, "The High/Scope Perry Preschool Program Cost-Benefit Analysis Using Data from the Age-40 Followup," *Journal of Human Resources* 41, no. 1 (2006): 162–190, www.jstor.org/stable/40057261; James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz, "The Rate of Return to the High Scope Perry Preschool Program," *Journal of Public Economics* 94, nos. 1–2 (2010): 114–128, https://www.nber.org/papers/w15471; and Lawrence J. Schweinhart, *Significant Benefits: The High/Scope Perry Preschool Study through Age 27. Monographs of the High/Scope Educational Research Foundation* (Ypsilanti, MI: High/Scope Educational Research Foundation, 1993).

9 As often happens with education research, subsequent studies have challenged and qualified the promise of ECE supplied by the earlier studies. Two recent experimental studies reported no differences in outcomes for treatment and control groups by the third grade, suggesting that effects may fade out over time. See Michael

Puma, Stephen Bell, Ronna Cook, Camilla Heid, Pam Broene, Frank Jenkins, Andrew Mashburn, and Jason Downer, *Third Grade Follow-Up to the Head Start Impact Study: Final Report,* OPRE Report 2012-45 (Washington: Administration for Children and Families, 2012), https://www.acf.hhs.gov/sites/default/files/opre/head_start_report_0.pdf; and Mark W. Lipsey, Dale C. Farran, and Kelley Durkin, "Effects of the Tennessee Prekindergarten Program on Children's Achievement and Behavior through Third Grade," *Early Childhood Research Quarterly* 45 (2018): 155–176, https://doi.org/10.1016/j.ecresq.2018.03.005. Other studies present varied results for program effects over time. See, for example, Carolyn J. Hill, William T. Gormley Jr., and Shirley Adelstein, "Do the Short-Term Effects of a High-Quality Preschool Program Persist?" *Early Childhood Research Quarterly* 32 (2015): 60–79, https://doi.org/10.1016/j.ecresq.2014.12.005. One conclusion from the contending studies is "to focus on understanding the conditions under which ECE programs yield positive and persistent results." See Daphna Bassok and Mimi Engel, "Early Childhood Education at Scale: Lessons from Research for Policy and Practice," *AERA Open* 5, no. 1 (2019): 1–7, https://doi.org/10.1016/j.ecresq.2018.03.005.

10 Robert Pianta, Jason Downer, and Bridget Hamre, "Quality in Early Childhood Education Classrooms: Definitions, Gaps, and Systems," *Future of Children* 26, no. 2 (2016): 11–29, https://files.eric.ed.gov/fulltext/EJ1118551.pdf.

11 Kathy Sylva, Iram Siraj-Blatchford, and Brenda Taggart, *ECERS-E: The Four Curricular Subscales Extension to the Early Childhood Environment Rating Scale (ECERS-R),* 4th Ed. (New York: Teachers College Press, 2010).

12 Virginia Vitiello, Daphna Bassok, Bridget K. Hamre, Daniel Player, and Amanda P. Williford, "Measuring the Quality of Teacher-Child Interactions at Scale: Comparing Research-Based and State Observation Approaches," *Early Childhood Research Quarterly* 44 (2018): 161–169, https://doi.org/10.1016/j.ecresq.2018.03.003. See also Terri Sabol, Sandra L. Soliday Hong, Robert Pianta, and Margaret Burchinal, "Can Rating Pre-K Programs Predict Children's Learning?" *Science* 341 (2013): 845–846, https://doi.org/10.1126/science.1233517.

13 For reviews of this evidence, see Bridget K. Hamre, Robert C. Pianta, Jason T. Downer, Jamie DeCoster, Andrew J. Mashburn, Stephanie M. Jones, Joshua L. Brown, Elise Cappella, Marc Atkins, Susan E. Rivers, Marc A. Brackett, and Aki Hamagami, "Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in Over 4,000 Classrooms," *Elementary School Journal* 113, no. 4 (Chicago: University of Chicago, 2013): 461–487, https://psycnet.apa.org/doi/10.1086/669616.

14 Robert Pianta, Karen La Paro, and Bridget Hamre, *Classroom Assessment Scoring System (CLASS) Manual, Pre-K* (Baltimore: Brookes Publishing, 2008).

15 Andrew J. Mashburn, Robert C. Pianta, Bridget K. Hamre, Jason T. Downer, Oscar A. Barbarin, Donna Bryant, Margaret Burchinal, Diane M. Early, and Carollee Howes, "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills," *Child Development* 79, no. 3 (2008): 732–749, https://doi.org/10.1111/j.1467-8624.2008.01154.x.

16 Effect sizes across studies for CLASS have been modest, prompting discussion in the field about the proper standard for evaluating the results; see Margaret Burchinal, "Measuring Early Care and Educational Quality," *Child Development Perspectives* 12, no. 1 (2018): 3–9, https://doi.org/10.1111/cdep.12260. Two points seem salient here. First, any in-school intervention can only be judged based on what is possible to achieve in a school setting, that is, affecting children from diverse backgrounds during a nine-month period, excluding the many out-of-school influences on learning. Second, modest increments accumulate over many years of schooling, adding up over time to produce more significant outcomes for children. From this perspective, CLASS effects are notable, even as effect sizes are typically small (< 0.1).

17 Notable is that while studies have revealed a relationship of CLASS to a range of outcomes, index-based measures have not fared as well. The active "ingredient" in QRIS systems that include CLASS is most likely the scores on CLASS that contribute to the QRIS index measure. See Sabol et al., "Rating Pre-K Programs."

18 Robert Pianta, personal communication, May 13, 2019.

19 Robert C. Pianta and Bridget K. Hamre, "Implementing Rigorous Observation of Teachers: Synchronizing Theory with Systems of Implementation and Support, in *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures,* eds. Jason Grissom and Peter Youngs (New York: Teachers College Press, 2016): 22–36.

20 The original CLASS system focused on early childhood education. Subsequently, CLASS has been modified and extended to apply to grade levels through 12th grade, with attendant studies exploring validity, reliability, and other issues. Martine L. Broekhuizen, Irina L. Mokrova, Margaret R. Burchinal, Patricia T. Garrett-Peters, and Family Life Project Key Investigators, "Classroom Quality at Pre-Kindergarten and Kindergarten and Children's Social Skills and Behavior Problems," *Early Childhood Research Quarterly* 36 (2016): 212–222, https://doi.org/10.1016/j.ecresq.2016.01.005; Margaret Burchinal, Nathan Vandergrift, Robert Pianta, and Andrew Mashburn,

"Threshold Analysis of Association between Child Care Quality and Child Outcomes for Low-Income Children in Pre-Kindergarten Programs," *Early Childhood Research Quarterly* 25, no. 2 (2010): 166–176, https://doi.org/10.1016/j.ecresq.2009.10.004; Bridget K. Hamre and Robert C. Pianta, "Can Instructional and Emotional Support in the First-Grade Classroom Make a Difference for Children at Risk of School Failure?" *Child Development* 76 (2005): 949–967, https://doi.org/10.1111/j.1467-8624.2005.00889.x; Mashburn et al., "Measures of Classroom Quality"; Eija Pakarinen, Noona Kiurua, Marja-Kristiina Lerkkanen, Anna-Maija Poikkeus, Timo Ahonen, and Jari-Erik Nurmi, "Instructional Support Predicts Children's Task Avoidance in Kindergarten," *Early Childhood Research Quarterly* 26 (2011): 376–386, https://doi.org/10.1016/j.ecresq.2010.11.003; and Sara E. Rimm-Kaufman, Tim W. Curby, Kevin J. Grimm, Lori Nathanson, and Laura L. Brock, "The Contribution of Children's Self-Regulation and Classroom Quality to Children's Adaptive Behaviors in the Kindergarten Classroom," *Developmental Psychology* 45 (2009): 958–972, https://psycnet.apa.org/doi/10.1037/a0015861.

21 The 2007 Head Start legislation provided requirements under the "Designation and Renewal System" that included the following criteria for renewal: "Any program that scores below the following 'floors' or 'low-quality thresholds' on any of the three CLASS domains would be required to compete: Instructional Support – below 2; Emotional Support – below 4; Classroom Organization – below 3. Note that the differences in these scores reflect research on the relationship between scores and child outcomes. Any program that scores in the bottom 10 percent on any of the three CLASS domains would be required to compete for continued funding, except that if a program scores in the bottom decile but the score equals or exceeds the exceptional level of quality threshold, the program would not be required to compete on the basis of this criterion." See https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/head-start-designation-renewal-system-final-rule.pdf.

22 There is an important point to concede here. Information from indicators may well provide a useful basis for improvement efforts, but these efforts still will need to be carefully designed, skillfully implemented, and scaled across many classrooms. The science supporting such improvement efforts includes many unanswered questions about optimal approaches. Indicators, in this sense, provide the starting point to what in any event is likely to be a long-term prospect, which policy scholars years ago referred to as "steady work." Richard F. Elmore and Milbrey Wallin McLaughlin, *Steady Work. Policy, Practice, and the Reform of American Education* (Santa Monica, CA: RAND Corp., 1988). For more on this point, see Bridget K. Hamre, Ann Partee, and Christina Mulcahy (2017): "Enhancing the Impact of Professional Development in the Context of Preschool Expansion," *AERA Open* 3, no. 4 (2017): 1–16, https://doi.org/10.1177%2F2332858417733686. For additional information, see Jason T. Downer, Robert C. Pianta, Xitao Fan, Bridget K. Hamre, Andrew Mashburn, and Laura Justice, "Effects of Web-Mediated Teacher Professional Development on the Language and Literacy Skills of Children Enrolled in Prekindergarten Programs," *NHSA Dialog: Research-to-Practice Journal for the Early Childhood Field* 14, no. 4 (2011): 189–212, https://doi.org/10.1080/15240754.2011.613129; Diane M. Early, Kelly L. Maxwell, Bentley D. Ponder, and Yi Pan, "Improving Teacher-Child Interactions: A Randomized Control Trial of Making the Most of Classroom Interactions and My Teaching Partner Professional Development Models," *Early Childhood Research Quarterly* 38 (2015): 57–70, https://doi.org/10.1016/j.ecresq.2016.08.005; Bridget K. Hamre, Robert C. Pianta, Margaret Burchinal, Samuel Field, Jennifer LoCasale-Crouch, Jason T. Downer, Carollee Howes, Karen La Paro, and Catherine Scott-Little, "A Course on Effective Teacher-Child Interactions: Effects on Teacher Beliefs, Knowledge, and Observed Practice," *American Educational Research Journal* 49, no. 1 (2012): 88–123, https://doi.org/10.3102%2F0002831211434596; and Robert Pianta, Bridget Hamre, Jason Downer, Margaret Burchinal, Amanda Williford, Jennifer LoCasale-Crouch, Carollee Howes, Karen La Paro, and Catherine Scott-Little, "Early Childhood Professional Development: Coaching and Coursework Effects on Indicators of Children's School Readiness," *Early Education and Development* 28, no. 8 (2017): 956–975, https://doi.org/10.1080/10409289.2017.1319783.

23 Robert Pianta, personal communication, February 11, 2019. More generally, a recent meta-analysis of quasi-experimental studies finds that professional development for early childhood educators is associated with measures of program quality and, through this pathway, to improvements in child development outcomes. See Franziska Egert, Ruben G. Fukkink, and Andrea G. Eckhardt, "Impact of In-Service Professional Development Programs for Early Childhood Teachers on Quality Ratings and Child Outcomes: A Meta-Analysis," *Review of Educational Research* 88, no. 3 (2018): 410–433, https://doi.org/10.3102%2F0034654317751918.

24 Center on Research and Evaluation (CORE), *Effects of Sustained Quality in PreKindergarten, Kindergarten, and First Grade in DallasISD: Executive Summary* (Dallas: Southern Methodist University, 2018), https://www.smu.edu/Simmons/Research/Center-On-Research-Evaluation/Sustained-Quality-in-Early-Grades.

25 Raters in Louisiana are selected and trained locally. One study has compared ratings by these local raters to those by a research team (Vitiello, "Teacher-Child Interactions"). Local raters gave higher scores on the Instructional Support domain and, in turn, the overall scores were somewhat higher and more variable. However, for both groups of raters, the total scores were associated with children's learning gains. This is an important finding in terms of scaling up observation systems that require human raters.

26 Abbie Lieberman, *Lessons from the Bayou State: Three Reforms for Improving Teaching and Caregiving* (Washington: America First, 2018): 20, https://www.newamerica.org/education-policy/reports/lessons-louisianas-early-childhood-system/.

27 Rachel Valentino, "Will Public Pre-K Really Close Achievement Gaps? Gaps in Prekindergarten Quality between Students and across States," *American Educational Research Journal* 55, no. 1 (2017): 79–116, https://doi.org/10.3102%2F0002831217732000.

28 Lynne Vernon-Feagans, Irina L. Mokrova, Robert C. Carr, Patricia T. Garrett-Peters, Margaret Burchinal, and Family Life Project Key Investigators, "Cumulative Years of Classroom Quality from Kindergarten to Third Grade: Prediction to Children's Third Grade Literacy Skills," *Early Childhood Research Quarterly* 47 (2019): 531–540, https://doi.org/10.1016/j.ecresq.2018.06.005.

29 Ibid.

30 Valentino, *Public Pre-K.*

31 David K. Cohen and Heather C. Hill, *Learning Policy. When State Education Reform Works* (New Haven: Yale University Press, 2001).

32 Dallas and Louisiana offer contrasting cases in this respect because CLASS operates in a low-stakes improvement context in the first case, and as a high-stakes accountability measure in the second case, respectively. Tracking the relative merits of each approach will be informative for future policymaking in this regard.

33 As noted above, CLASS has been revised and extended to operate across grade levels in K–12 education, with similarly promising results as an indicator of quality. In principle, then, CLASS could serve as an indicator of quality across the entire education system, notwithstanding the political obstacles as discussed.

34 An update to this observation: Louisiana now has such information covering four years for all toddler and pre-K, publicly funded programs. Robert Pianta and Bridget Hamre, "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity," *Educational Researcher* 38, no. 2 (2009): 109–119, https://doi.org/10.3102%2F0013189X09332374.

Measuring the Power of Learning.®

www.ets.org