District
Management
Group

Accountability That Improves: An Interview with Jim Liebman

> *It is good to know whether you did well or not. But it is much, much more important to know why.*

Accountability That Improves:

An Interview with Jim Liebman

From 2006 to 2009, Jim Liebman built and operated a comprehensive performance management system in our nation's largest public school district. As the first Chief Accountability Officer at the New York City Department of Education (NYCDOE), Liebman effected an array of reforms, including the design and implementation of A to F letter grades for each school, the development of an $80 million data system to track student performance in minute detail, and a citywide scale-up of school-based collaborative inquiry teams. Although Liebman faced criticism from educators and parents along the way, the accountability and support tools he put in place have changed the way many think about public education and have become models for reform-minded educators nationwide.

Liebman's redesign of the way student outcomes were measured and managed came at a time of dramatic change for New York City's public schools. Chancellor Joel Klein was implementing a number of reforms based on a theory of action that prized principal empowerment. Schools were given greater levels of autonomy, the organization was restructured, and a new budgeting system titled "Fair Student Funding" was implemented. Through these reforms, Liebman and Klein hoped to spur innovation and increase educational performance.

Before moving to the NYCDOE, Liebman had built a distinguished career as a public interest lawyer and a law professor. He has spent most of his career writing and teaching about inequality in school and criminal court systems. Since leaving the NYCDOE, Liebman has returned to Columbia Law School where he directs the Center for Public Research and Leadership, which immerses students from a variety of professional schools in public education improvement strategies.

In this edited interview, DMC CEO John J-H Kim, Associate Daniel Goldberg, and Professor Jim Liebman discuss ways that thoughtful measurement of student outcomes can improve educational performance in a district.
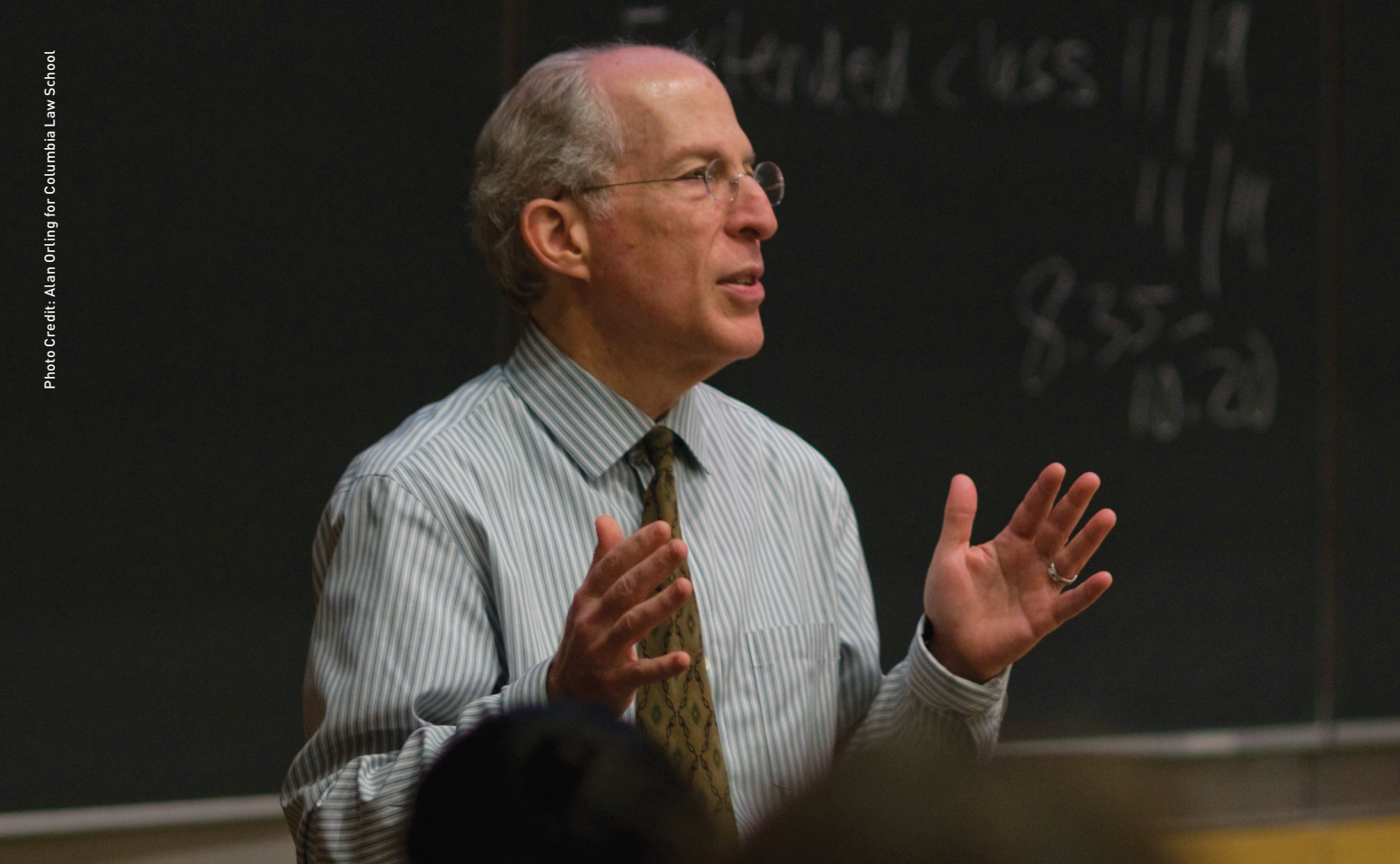
Jim Liebman at Columbia Law School

You spent most of your career as a law professor. How did you get involved in New York public schools?

The area in which I practiced law was civil rights, and a large part of my work had to do with school desegregation. Ever since then, I have been interested in the mechanics of public schools—the ways that schools can be organized and influenced to achieve their mission to successfully educate students across the spectrum.

Gradually, with the standards movement starting to take hold and other kinds of institutionally focused reforms happening in the United States, my work began to shift away from civil rights interventions in court to thinking more about more structural and organizational interventions. In writing about school desegregation, I developed an interpretation of school desegregation as a political and an institutional mechanism for reform, rather than a way of simply rearranging the bodies within schools and districts. What I call "all-out desegregation" forces districts and schools to change in dramatic ways and to take seriously the needs of poor and mi-

> " **It's good to know whether you did well or not. But it's much, much more important to know why.** "

nority children by putting them in a position where their voices would have to be considered along with others'. I was influenced by John Dewey, and thought a lot about how, instead of simply ordering organizations to do the right thing, you can restructure them to build into their everyday thinking and action the motivation and the capacity to serve those most in need. I believe that is the only way to get organizations like school systems to really support poor and minority children. →

> **"Each school should aim to be the best in the city, but in the meantime, you should strive to do as well as, or better than, the best schools that have students like your own."**

Before I came to Columbia Law School, I had litigated the Kansas City school desegregation case against Joel Klein—I was representing the black children in Kansas City, and he was representing the state of Missouri. When he became chancellor here, I started emailing him about my kids' school, just like so many other parents in the city. He remembered me, and we struck up a little bit more of a relationship. Eventually, he asked me if I might want to come work for him.

Did you win the case or did Joel Klein win?

Well, he won it at the trial level. We took it up on appeal, and the case was decided by an even vote of 4-4 because the ninth judge got sick on the day of the argument. That meant he won, because a tie goes to whoever won below, but I still remind him that we tied.

It seems to me that the first thing you have to do in establishing an accountability system is get an agreement or a definition of what educational outcome you're shooting for. Do you think that is true?

I think that's a really important point. My view is that you want to give educators as much freedom as you can about how to get to the point you want to get to, but it's very important that there be

some democratic reckoning about what that point is—about the outcome you are aiming for. I am a believer that you can't leave that question up to individual educators to decide for themselves and for whichever children happen to end up in their classroom.

The "what" has to emerge from some sort of democratic process that identifies what the state or district as a whole feels are its priorities—something that the switch to mayoral control finally permitted New York City to do as of 2002. There had been a strong consensus building across the country in the 1980s through the 1990s—which continues today, though some are challenging it now—that demonstrable academic outcomes, which people know are highly correlated with graduation from high school and success thereafter, are the key goal of school districts. We're now enriching that by thinking about some of the social and emotional and other kinds of background skills that are needed, but I would say that the combination of those have typically emerged from consideration at the district or state level as to what the outcomes should be. I would define those outcomes as academic success within the K-12 system that leads to graduation from high school with the capacity to succeed in college and career. And it was those objectives that Mayor Bloomberg embraced on behalf of the voters of the city who elected him and charged Chancellor Klein with accomplishing.

How did you decide which metric is best for any particular outcome?

I think about this in much the same way as the people who developed the balanced scorecard and other evaluation systems. You need to identify your goal, and then your initial theory of action for getting there. If your goal is to accelerate student learning, what do you think needs to happen in order for the district to enable schools, and schools to enable educators, and educators to enable students? Then you need to decide how to measure whether each of those steps is occurring and having its desired effect. And if it is not, you need to

use a structured inquiry to figure out whether it's the theory of action, the implementation, or the metric that isn't working right, and adjust.

One question is: how do you measure whether educators and students actually accomplish the ultimate outcome? But just as important as those "lagging indicators," because it really covers many more years and many more activities, is the question of how you identify whether the intermediate steps that you think are crucial along the way are in place. We used things like qualitative reviews and surveys as leading measures. Without them, you can't know that you're doing well what you need to do in order to succeed.

What makes one of these metrics better than another?

It's very important that these measures have a couple of features. One is that they not ask people to accomplish anything that they cannot realistically accomplish. The metric should demonstrate to them—through benchmarks and targets—that the desired outcome is within their capacity to achieve. One way to accomplish that is to base your metrics on evidence of how the highest-performing comparable schools, comparable educators, and comparable classrooms have done in the recent past. Another is to benchmark against the past performance of the school itself or the educator him– or herself, and say, "You need to do as well as you've ever done in the past and a little bit better." But it always needs to be based on proven experience; educators will ignore a request to meet a metric that there's no evidence they can meet, as they did when first faced with No Child Left Behind. Why should educators take that seriously?

And then the second important feature is that the metric be as diagnostic as possible. I think of this overall process as maybe one part evaluation and three parts diagnosis. It's good to know whether you did well or not. But it's much, much more important to know why. That means knowing how individual children did overall and on individual components of the skills and knowledge they need, so educators can see where they've succeeded and where they've failed.

What's an example of a good metric, and how did you design it?

When you're asking educators to move children academically, it's important to measure whether the students have gotten to where you want them to arrive. Yet, in many classrooms, educators are working with students who need to catch up and are not going to get there by the end of that school year or even within a few school years. So, if you just measure students by the static outcomes you want—some proficiency level, or whatever—you're asking, in many cases, too much. Instead, you want to measure longitudinal growth: how much each child, on average, has gained from the beginning to the end of the year.

But when you do that, you need to measure it in a way that the educator can understand. I am not a fan of regression models for measuring growth, which compare all this stuff in some kind of algorithmic soup that nobody can understand. Instead, what we tried to do in New York City was to measure growth using metrics that any teacher using plain old arithmetic could reverse-engineer him– or herself, and therefore be able to say, "Here's where the student was at the beginning of the year, here's where the student is at the end of the year, and here's how it affected my overall result. I can make →

the same calculation the district made to tell me how well I did with that student, and I can see where the success was and where more work is needed." The data needs to be very transparent and allow a lot of diagnosis so you can see that some kids moved a little, others moved a lot, and which particular skills are involved in that learning.

Did you do any segmentation of students' growth scores?

Yes, we absolutely did. We really felt that it was important not to base everything on a comparison of a school that had all low-performing students to a school that had all high-performing students. That's sort of saying, "You should be able to accomplish magic." That's not a measure of what educators add to kids' learning; that's a measure of what kids brought the first day. And that's not fair to educators.

So how did you compare schools instead?

We came up with what we called the "peer rating" for every school in the district. It was a measure of the level of challenge that kids in that school present, and it was substantially based on incoming test scores for middle and high schools. For elementary schools, we had to use more demographic variables—the percentage of students receiving free and reduced lunch, percentage of English Language Learners, percentage of special education students, and also percentage of African

American and Latino students—because we know, statistically, these populations can present some additional challenges.

For every metric, we compared each school to the 40 schools in New York City most like it. There were 1,600 schools, so we had a lot to work with. Each school was always in the middle of its group of 40 schools. There were always 20 schools that were slightly higher on the peer ranking (that is, facing more challenges) and 20 slightly lower. Then at the beginning of the year, we told every school, "Here's your range: here's how the lowest-performing schools in your cohort did on this metric over the past three years, all the way to how the highest performing of your peer schools did in that period. You will be measured by how close you get to how the top-performing school in your cohort did on that metric. And if you do even better, we'll give you credit for that, and going forward, your performance will be the benchmark for other schools like yours." So we were using a real "race to the top" approach.

That made up about 75% of the grade for each metric, and then 25% was based on how your school compared to all schools of your type—elementary, middle, or high school—in the city. The theory was that each school should aim to be the best in the city, but in the meantime, you should strive to do as well as, or better than, the best schools that have students like your own. This motivated educators, because they believed we were holding them to a standard they could meet because schools like theirs had met it in the past.

Jim Liebman's Biographical Timeline

1974

B.A., Yale University

1977

J.D. Stanford Law School

1978

Law clerk to Justice John Paul Stevens

1979

Assistant Legal Counsel, NAACP Legal Defense and Educational Fund

Do you think it makes sense to be doing that measurement at the school level, or do you think it makes more sense, if possible, to do it at the student level?

Well, I actually think you need to measure and make these kinds of comparisons for every student—as well as every classroom, every district, every state. We reported student information to every school using a data system that we created. It is important to have information at all of these levels for diagnostic purposes. But the question becomes, at what level will you be aggregating those results to reach a summative judgment, for stakes? And my strong view after thinking about this a lot—and this is controversial—is you should do that at the school level and above. You need to report the data at all levels, but high-stakes judgments below the level of the school are harder because you get into problems of reliability and predictability.

But won't you get the pushback that change really happens at the classroom level? Shouldn't we be holding teachers accountable and not the school?

I understand that argument. I certainly would calculate everything and make comparisons at the teacher level, as well as at the school level, and I would make that visible to the teacher and make it visible to the principal and the district. But I would say, "Principal, we're holding you accountable for your school. You are in charge of doing everything you can, including working with your teachers and bringing them into the process of deciding how you're going to succeed at the school level." Any principal knows that the most important ingredient for the school's success is the teachers.

> **" We've turned measurement into a technical question rather than a political or ideological question. Once it's a technical question, we can get better at it. "**

In other words, I believe in holding principals accountable for the improvement of their teachers and for the teachers' outcomes. Then it's up to the principal, motivated by the school's accountability, to decide what to do. If there are teachers who after several years have not improved from an ineffective level, I would ask the principal to give the district superintendent an explanation for that. But the minute you start specifying for all schools across an entire state exactly what needs to happen after one or two years of ineffectiveness, that's too inflexible. Principals are going to have a much better sense of what they need to do with each educator, and as long as the principal feels responsible for the school's outcomes, the principal will have the motivation to figure out how to move those teachers forward or potentially move them out of the system. I just wouldn't specify that, for every school, principal, and teacher, there's one right answer. ➔

1985
Joined Columbia Law School faculty

2002
Champions of Justice Award, National Association of Criminal Defense Lawyers

2006
Chief Accountability Officer at NYCDOE

2009
Returned to Columbia Law School

Jim Liebman—the first Chief Accountability Officer at the New York City Department of Education (2006–2009)

Do you think there's a difference in the data that you need to drive accountability versus data that informs practice? The conversation about accountability is pretty high stakes, but if you just wanted to improve practice, should the type of data tracked be different?

I am a big believer in getting the right balance of what you might call external accountability and internal accountability. External is when the system imposes metrics; if you don't do well on those metrics, there are consequences and, if you do well, there are rewards. I believe there is a place for that approach to create motivation and to get everyone in the system to understand who really matters and how important certain outcomes are—the kids matter; their learning matters.

But I also think that it's really important to get internal accountability, where teams of educators, particularly at the school level, feel mutually responsible for the kids at their school and hold one another responsible for the outcomes. To get there, you've got to have trust, and in order to have trust, you have to have metrics and data you are willing to rely on. To have faith in that data, you need to be willing to allow the data to be questioned and examined by groups of teachers in a strategic and structured inquiry.

How does one get "the right balance" between these two kinds of accountability?

It's about context. Some school systems already are widely committed to kids' success, and they can focus on deepening internal accountability. But other districts have fallen into a process over the years of believing that a lot of kids can't learn—maybe because of their parents, their neighborhoods, their attitudes—and therefore that it's the adults' interests and comfort that matter the most. Those districts may need more external accountability at the start.

In either case, you need some metrics that are diagnostically important, but that don't have stakes attached to them. As I said, I do think it's important to have stakes at the school level, but at the teacher level, I think stakes can impede internal accountability—the creation of a degree of trust and transparency that can lead educators to be willing to talk to each other, share their outcomes, and be frank with each other about what needs to change. That takes some real work and trust, and high stakes can be an impediment.

Generally, do you think teachers and schools have the right information about practice, outcomes, and cost that will lead to a cycle of continuous improvement?

Typically, the answer is no. But it's very difficult to guarantee that information unless you create good inquiry processes within schools where teachers start to explore these questions themselves and generate a demand for the kinds of information they need. Then I think you start to get a better sense of what information is needed. There definitely is information that everybody's going to need, but there's a lot more that's specific to particular children and classrooms. To get that information, teams of teachers need to be involved in strategic and structured inquiry, asking questions and then identifying where they need more data and don't have it. Then they and their district facilitators can come up with strategies for getting the information, and teachers can put what they learn into effect, measure if it's working, and adjust.

What would you say has changed the most in how we measure outcomes over the past decade?

I think the biggest change is almost at the conceptual level. It used to be that we thought we couldn't measure anything unless we could figure out a perfect way to do it beforehand. I think the biggest change has been to say that some information is better than none. Then, once we start measuring information and put it out there with some importance on it, you start having people weigh in about what's wrong with the data, and it is at that point that the data can and will improve.

The most important thing that's happened over the last ten years is that we got started. We started with some stuff that wasn't very good, like No Child Left Behind. But even that motivated educators in better directions, and since then, there's been a real effort to improve the quality of the metrics. And we've seen dramatic improvement in the ways we measure growth, and in the way we compare schools to one another. We've turned measurement into a technical question rather than a political or ideological question. Once it's a technical question, we can get better at it.

If you could do anything in terms of strengthening accountability and practice, what would you want people to study or do?

I would say two things. One is that we should be attentive to enriching qualitative analysis—both qualitative review of schools and qualitative observation and evaluation of teachers—with the same degree of intensity that we have applied to more quantitative metrics over the past five to ten years. I'm not saying we should replace a reliance on quantitative metrics, but we should enrich it further with the qualitative. I think there is a rich capacity for improvement in rigorous qualitative reviews that remains to be unlocked.

And second, we've gotten much more sophisticated at the district, school, and principal levels at understanding measures—their value, their rigor, what they can tell us diagnostically. But I think that we need to spend a lot more time and focus bringing teachers into that process in a rigorous and genu-

ine way. Teachers need to be part of the leadership and the thinking and the implementation of this. When we get the value of measurement down to the classroom level and the teacher level, then the next big burst of improvement is going to happen.

> " **When we get the value of measurement down to the classroom level and the teacher level, then the next big burst of improvement is going to happen.** "

With a new mayor and a new chancellor in place in New York City, it seems that some of the ideas and the work that you've implemented may now be changed significantly, and in some cases eliminated. What's your sense about that?

There is not one right way to do this; there are many ways to do it. There's internal accountability, external accountability, and there's a balance. If they want to restrike that balance in favor of internal accountability, that's fine. I just hope that there won't be a retreat from the basic agreement, which I think has been reached in this country over the past decade or two, that it's more important than anything else for educators to feel responsible for the academic outcomes of their students and to feel capable of affecting and improving those outcomes.