National Council on Measurement in Education. Inaugural Opening Plenary Session. April 14, 2012, 10:45 a.m., Hyatt Regency Ballroom A.

John Q. Easton

Thank you for the kind invitation to speak to you today. As at best a "dabbler" in psychometrics and measurement, I am keenly aware that you folks have formidable technical skills and experiences that I lack. So I hope that you will indulge me a bit as I talk about several issues that I know you know more about than I do.

I want to break this talk into three sections, which I hope will have some semblance of cohesion when I'm done.

- First I want to talk about IES. You probably know pretty much what we do and who we are, but I am going to speak about the specific goals that I hope to accomplish in my six-year term as director. I will also talk about how my previous experiences in Chicago shaped these goals.
- Next, I want to talk a little about "the power measurement," an idea that you don't need to be convinced of.
- Third, I want to tell you my priorities on how I think that the assessment and measurement community can help promote and engage in school improvement today.

Part One. Some background on IES.

IES was established in 2002 by the Education Science Reform Act. Built from the Office of Education Research and Improvement, IES now includes four centers. The National Center for Education Research and the National Center for Special Education Research both primarily make grants for research and training, many of you know that we have sponsored considerable measurement –related research; in fact we have a research goal on measurement. The National Center for Education Evaluation and Regional Assistance evaluates federal programs and sponsors the Regional Education Laboratories, the What Works Clearinghouse, and ERIC, the Education Resources Information Clearinghouse. Fourth, the National Center for Education Statistics, which conducts the National Assessment of Educational Progress, international assessments, including PISA, TIMSS, PIRLS, and PIAAC, and collects scores of other statistics, including our premier longitudinal surveys, such as the Early Childhood Longitudinal Study (ECLS), and the High School Longitudinal Study of 2009.

Many of you here know more about NAEP than I do. I see people who serve or have served on the National Assessment Governing Board and helped set NAEP policy; others here are on technical advisory groups, like the NAEP Validity Panel and the Design Advisory Committee; others of you actually create, administer, and score NAEP and help report the results. What we all do know

about NAEP is that it's not a single test or even a series of tests, but a very, very complex system, with many more moving parts than most people could possibly imagine. NAEP is at a bit of a cross roads right now, facing many big questions and challenges. I will come back to NAEP later.

IES is the government and the country's agency for education research, evaluation, statistics and assessment. Above all, we strive to be accurate and objective. Because we must be seen as non-partisan, non-political and trustworthy, we have some independence from the Department of Education. For example, we have our own review process for both funding decisions and for our publications. At the same time though, we must strive to provide policy relevant information in a timely fashion to decision makers in the government and elsewhere. The IES FY12 budget appropriation is just under $600 million, and the President has requested an increase to $621 in FY13.

I've been at IES for almost three years now and my singular goal is to make our work as relevant and usable to both policy makers and practitioners as possible. I talk about this goal for our work at IES at every opportunity. I think that our research and evaluation, and let me add, our measurement, can be more relevant and usable to practitioners and policy makers. By striving to be relevant and useful I also believe that we are in a stronger position to build a more robust science of education that will help us understand more about the school

improvement process, what constitutes better teaching, and how to support more student learning and the policies and practices that we need to put in place to reach these goals. As we often say and hear, education research needs to move beyond trying to discover "what works" to learning about why, when, where, for whom and under what conditions practices, programs and policies work.

IES is noted for rigor, thanks to the efforts of its first director and my predecessor, Russ Whitehurst, who really raised the quality standards and expectations for education research across the country. We are now at the point where we can aim for that sweet spot where we conduct and sponsor research that is both relevant and rigorous. I am intent on retaining rigor for which IES has made its mark, but also intent on expanding our efforts from a distinct focus on validating programs, interventions, etc. to creating a broader view of building understanding of the systems, and context and the messiness and complications of school improvement and the outcomes that we value for our children. I think that the measurement community has a vital role to play in this work.

My thinking is clearly influenced by my work in Chicago, where I spent about 30 years working with or for the Chicago Public Schools doing research and providing information that would help guide improvement strategies. I really loved that work and believe that my colleagues and I made many useful

contributions. I liked doing the research and the inherent challenges in making sense of a jumble of data, but I also got as much satisfaction from interacting with school leaders, discussing the implications of the work with them and especially learning from their perspectives and experiences.

I believe that partnerships between researchers and policy makers and practitioners are essential for moving the science of education forward, beyond the search for "what works." Partnerships help researchers focus more on "problems of practice" than on "problems of interest." Practitioners and policy makers are more likely to act on research findings when they've had a role in planning the studies and interpreting and making sense of the findings. I also think that the partnerships can help advance science, as policy makers and practitioners push researchers to take next steps together, help figure out what to do, try it, study it and continue to learn from it.

IES is asking more of its researchers to work collaboratively with practitioners. Our single biggest research program, Reading for Understanding, requires that the research teams work with practitioners as they develop new interventions to help K to 12 students comprehend text more deeply. Our research program, Evaluating State and Local Programs and Policies, requires partnerships between researchers and SEAs or LEAs. Our Regional Education Laboratories are now required to create research alliances with a broad range of

stakeholders to conduct research on high priority topics. These alliances are additionally charged with building a coherent series of research studies focused on the same topic. Finally, we just released a brand new Request for Applications, called "Researcher-Practitioner Partnerships in Education Research" that will provide modest funds for state and local agencies and researchers to work together on problems of practice or policy.

Part Two. The Power of Measurement.

Now, I'm switching to a different topic, where I'm perhaps especially vulnerable, given that I'm talking to people who know so much more about this than I do and it's much like "preaching to the choir." The title itself, "the power of measurement" is somewhat presumptuous. I've used some of this material in a talk before, unfortunately not with a lot of success, but I'm going to plunge ahead anyway.

The previous occasion was about 4 and a half years ago in Chicago when my organization, The Consortium on Chicago School Research at the University of Chicago, celebrated its 15th anniversary. We wanted to mark this occasion with an event that would serve multiple goals – it was meant to be fun and congratulatory but we also wanted it to serve an educative function and be reflective of both the content and the process of the Consortium's research and its public informing activities. Tony Bryk had left the Consortium for Stanford by

then, but he came back to help us with both the intellectual and the social content of the day. He gave an outstanding intellectually charged talk on pressing issues in urban education and promising approaches to solve these problems. Arne Duncan and Barbara Eason Watkins spoke about the Consortium's impact on the Chicago Public School system. My colleagues Elaine Allensworth, Stuart Luppescu, Melissa Roderick and Penny Sebring described the Consortium's research in lively and provocative presentations. And then I spoke about the "power of measurement." It wasn't exactly the most popular session of the day. I hope that you will be more appreciative.

We know that good measurement is the cornerstone of all scientific research.[1] I believe it is especially important in a social science like education that depends heavily on the interactions and communication among and between key stakeholders, whether they are parents, students, teachers, principals, researchers, superintendents, school board members or newspaper reporters. Good measurement brings conceptual clarity by precisely defining the phenomena that we are trying to change. It enables researchers to build frameworks or theories that integrate multiple concepts. It helps us better test and then understand the mechanisms and pathways to improved outcomes.

[1] The remainder of this section of this talk draws heavily on this report: Roderick, M., Easton, J.Q., & Sebring, P.B. 2009. *Consortium on Chicago School Research: A New Model for the Role of Research in Supporting Urban School Reform.* Chicago: Consortium on Chicago School Research. See pages 23-24.

Good measurement also gives educators frameworks to help them place the phenomenon in context as they plan or seek improvement strategies.

We can measure a lot of important phenomena in education, not just student achievement outcomes. We also measure more proximal outcomes and process variables that lead the way to school improvement, and psychological constructs about students and teachers that may be inputs, process variables or outcomes. Let me give a couple examples.

Elaine Allensworth and I wrote two papers on what we and our colleagues in Chicago called the "on track to graduate" indicator.[2] This is a simple binary indicator created at the end of the freshman year based on the number of course credits earned and the number of F's in major subjects.[3] A student is either on track to graduate or not. It turns out that this on-track indicator is a very good predictor of high school graduation. Students who are on-track are 4 to 5 times more likely to graduate than students who are off track. What's especially important here is that the on-track indicator is a better predictor of high school graduation than students' demographic characteristics, including their age, prior grade retention, school mobility history and their 8th grade test scores.

[2] Allensworth,E. & Easton, J.Q. 2005. *The On-Track Indicator as a Predictor of High School Graduation*. Chicago: Consortium on Chicago School Research; Allensworth, E. & Easton, J.Q. 2007. *What Matters for Staying On-Track and Graduating in Chicago Public Schools*. Chicago: Consortium on Chicago School Research.

[3] A student is on track if he or she has earned five full credits and no more than one semester F in a major subject by the end of the freshman year.

We did a lot of digging around to figure out why this indicator was so important. How are student behaviors (e.g., attendance, homework, or classroom engagement) associated with increasing the likelihood of students being on-track for graduation? Are the predictors of being on track freshman year different than the predictors of high GPA? And, what elements of school and classroom environments are associated with student behaviors that lead to improved class performance?

So here we have an indicator, or call it a variable or measure, that is reliable, has predictive validity and is simple. Good measurement brings about conceptual clarity enabling us to build and test models for the linkages among being on-track and school attendance, classroom engagement, doing homework and other relevant behaviors. But I want to stress how much good measurement facilitates communication among researchers and the wide range of stakeholders who are seeking ways to improve student high school graduation rates. A common vocabulary facilitates communication across stakeholder roles.

I want to give a second example of the power of measurement from our work in Chicago. This is an example of defining, testing, and measuring a construct critical to organizing schools for improvement and for helping educators in developing broader frameworks for what matters for school improvement. In one of the Consortium's first reports, *A View from the*

Elementary School,[4] Tony Bryk and several of our colleagues described a "Christmas tree" school. In this school, the principal used the resources provided by state funds to purchase an indiscriminate range of programs. These programs were compared to the ornaments on a Christmas tree, displaying a great deal of energy and innovative spirit. The problem was that all these new programs were unconnected and uncoordinated; teachers and students alike were adversely affected by this incoherence. In some instances, students moved from a whole-language approach to teaching reading in one grade to a direct instruction approach in the next grade. The study contrasted the Christmas tree school with a school where the programs were coordinated and aligned both across and within grades—the kind of practices that we called program coherence.

Our next step was to measure this concept through large-scale survey data collection. In what have now become bi-annual surveys of students, teachers and principals, we asked teachers a series of questions about the degree to which they feel: programs at their school are coordinated with each other and with the school's mission; instructional materials are consistent within and across grades; and there is sustained attention to quality program implementation. Together these items form a highly reliable scale. This survey measure then was validated

[4] Bryk, A.S., Easton, J.Q., Kerbow, D., Rollow, S.G. & Sebring, P.B. 1993. *A View from the Elementary Schools: The State of Reform in Chicago*. Chicago: Consortium on Chicago School Research.

by field work in the Chicago Annenberg Research Project.[5] Researchers independently and blindly rated a sample of schools on the degree of program coherence based on numerous visits, observations, and interviews. They found a high degree of correspondence between their own ratings and the survey results, providing statistical validity for the measurement scale.

Program coherence sounds like a good idea on face value, but most importantly, we have found that schools with high program coherence are more likely to improve student achievement over time; and, similarly, schools that become more coherent over time are also more likely to improve student achievement.[6] Here's a concept discovered and described in one study, then reliably measured though large scale surveys and validated in a field study. Finally, by linking to student achievement gains, "program coherence" becomes a leading indicator that can be tracked over time. And because it is such a cogent idea, all sorts of people can understand its meaning and grasp its importance and, most of all, take actions to increase program coherence in their schools and monitor their school improvement activities through this lens.

Good measurement alone isn't the sole answer to all of our school improvement needs, but it is a crucial and fundamental component of any serious research and improvement endeavor. Good measurement helps us

[5] Newmann, F.M., Bryk, A.S. & Nagaoka, J. 2001. *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Chicago: Consortium on Chicago School Research.
[6] Bryk, A.S., Sebring, P.B., Allensworth, E., Luppescu, S., & Easton, J.Q. 2010. *Organizing Schools for Improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

identify, define and clarify the nature of a problem. It can help us to determine the linkages among other behaviors and conditions. It can help us communicate across stakeholder communities. Measurement isn't just about outcomes, but also about the processes that we need to improve them. Reliable and valid process and leading indicators can help us develop and track the progress of improvement strategies.

Part Three. How can the measurement and assessment community further school improvement?

Right now, I could probably list a couple of dozen topics that I think are front-burner issues for the testing, assessment and measurement community. This entire conference is built around those topics and you will be talking about them for the next several days. I would like to give this my personal twist and speak to a number of topics that I believe are of special importance.

First, and probably most obvious, is that we need to dramatically improve the quality and depth of our large-scale state assessments and probe more deeply into students' higher order thinking and problem solving skills. The federal government, through Race to the Top Assessment, is investing a large sum of money in this effort and is hopeful for the results, but like all of us is anxious about the timelines, the ambitiousness and the technical demands of this project. We at IES are investing over 15 million dollars in a new low stakes

reading comprehension assessment system for K-12 as part of our Reading for Understanding initiative. (You can hear more about this at 4:05 on Sunday afternoon.) The Gordon Commission is working on this topic and will be discussing it later today.[7]

We have ample evidence to show that the nature and quality of high stakes assessments influence teachers' instructional practices[8] so there is widespread agreement now that we need higher quality and more demanding assessments for our students as part of an instructional reform agenda.

There's another argument in favor of the need for deeper assessments that is less well understood or thought through, at least by me. For decades now, researchers have consistently found weaker school and teacher effects on reading than on mathematics.[9] This is usually been explained by family and community influence over early oral language development, vocabulary acquisition, etc. Tom Kane and his colleagues replicated this finding in one of the Gates Foundation sponsored Measures of Effective Teaching project studies, but with a difference.[10] Teacher effects are smaller in language arts than in math, but only on the state mandate NCLB assessments. Teacher effects are just as large

[7] For more information on the Gordon Commission: http://www.gordoncommission.org/.
[8] See for example: Koretz. D. 2008. *Measuring Up*. Cambridge: The President and Fellows of Harvard College.
[9] Nye, B., Konstantopoulos, S., & Hedges, L. V. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.
[10] Kane, T.J. & Staiger, D.O. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle: Bill & Melinda Gates Foundation.

on a separate writing assessment as they are on the math assessment. Tom attributes the lower reading effects to the assessments themselves rather than the phenomenon of developing a range of language arts skills. Teachers do teach children complex and difficult language arts skills but our traditional assessments don't pick them up.

Second, I cannot miss the opportunity to talk about everyone's favorite subject: teacher evaluation. This is one of the single most difficult issues in front of us now. You can't escape it – barely a day goes by without a newspaper article, a lawsuit or a confrontation. It's an extremely tricky issue: obviously fraught with many hard technical questions that many of you here have thought long and hard about. But maybe even trickier are some of the ethical issues. Should teacher evaluations be part of the public record? If not public, what should parents have the right to know more about their children's teachers? How do we balance the need for transparency on the one hand and privacy on the other? Perhaps the most important question for this group is how do we best use imperfect information? We don't want to make consequential decisions based on faulty data and faulty analysis, but neither do we want to make the perfect the enemy of the good.

This is a topic where I think we in the research and measurement communities should throw ourselves right into the fray. All across the country

now states and local school districts are developing and implementing teacher evaluation systems that include some measure of student achievement or growth. They need people with our skills and expertise to help keep them out of trouble and avoid real damage. But we can do much more than that for them. We can help them develop and systematically and rigorously test teacher evaluation systems that can be helpful, productive, salutary. Good teacher evaluation can lead to better teaching and more student learning. It can lead to better decision making from administrators. It can lead to better teacher training, induction and professional development.

Teacher evaluation is a broad and varied topic, even though much of the discussion has focused on the controversies around value add measures of student achievement. But a good teacher evaluation system will have multiple components, combined in different ways for different purposes and perhaps at different times in teachers' careers. So aside from the value add, student achievement component, what are some of the others that are being used or should be used? Classroom observation, by principals, peers, and staff external to the school. Student survey reports of their classroom engagement. Principal holistic ratings. Teacher self assessments. Parent ratings. Scores on the intellectual quality of student work. Student social and emotional learning, behavior, collaboration. I think that there are probably dozens of potential

measures that could contribute to a robust teacher evaluation system. My request to you is to get involved in this on-the-ground work. Help a school, a district or a state develop, test and evaluate their teacher evaluation systems. This is a crying need. We have the skills to help do teacher evaluation better and we should all be stepping up.

Third, I want to talk briefly about the future of NAEP. Fortunately, there are a lot of very capable people thinking about how to keep NAEP as our "gold standard" assessment as the world of standards and assessments is rapidly changing. Ed Haertel is chairing a group that is preparing a report on this topic. Ed reported on their progress at a public meeting of the National Assessment Governing in March. One goal is to keep NAEP as the "backbone" of the student assessment infrastructure in the United States, by linking to international assessment and the new common core assessments. NAEP has served as a real engine of innovation in assessment and Ed's committee and we at IES and NCES would like to maintain this lead as an innovation laboratory. I look forward to the complete report and hearing the committee's recommendations on NAEP research and development agenda, the development of new item types, the interpretability of scores, technology enhanced accommodations, revisiting of achievement levels and scores of other topics. I'm sure this agenda,

supplemented by other ideas, will require major investments of human capital and other resources from the measurement and research communities.

I've mentioned three major critical needs so far: better large-scale assessment, better teacher evaluation systems, and the future of NAEP. There are probably dozens of other high priorities topics that I could have chosen: improving formative and curriculum-embedded assessment; computer based and computer adapted testing; more efficient scoring of open ended, extended and constructed response items through automated scoring; assessing 21st century skills like collaborative problem solving; and assessing college and career ready skills. There are countless others.

But I want to conclude by moving in a different direction. Let me step back and describe a recent study that's received a great deal of press coverage and attention, called "The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood,"[11] by Raj Chetty and John Freidman from Harvard and Jonah Rockoff from Columbia. These researchers assembled an enormous data set that allowed them to follow over 2.5 million children from grades 3 to 8 into early adulthood. They matched school test score and administrative data with tax records for the same students 10 to 12 years later so that they were able to study the relationship between students' school

[11] Chetty, R., Friedman, J.N. & Rockoff, J.E. 2011. The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. National Bureau of Economic Research. Working Paper 17699.

experiences and their adulthood success factors like earnings, college attendance and teenage births.

Chetty and his colleagues found that students who had one high value add teacher (+1SD above average) in reading or math in grades 4-8 were more likely to attend college at age 20, have steeper earning trajectories and reduced likelihood of having children as a teenager. This study provides additional clear evidence that good teachers (measured narrowly by value-add) matter, not just in the short-term, but in the longer run as well.

What I find fascinating about this study is that the high value add teachers gave students a bump up in their test scores and in their learning rates, but 2/3 of the bump faded out after a few years. As you know, this is a common finding in education research. An intervention that may look good in the short term loses its impact over the long term. But here we have the fade out in test scores yet we still see the improvements in very important distal outcomes, like going to college, earning money and avoiding teenage births.

What's going on here? I hypothesize that the effective teachers, measured by high value adds, are indeed boosting their students' achievement, but they are also boosting other important skills, traits or attributes that aren't measured in this study. I am betting that they could be psychological constructs like grit, perseverance, self-control, engagement, emotional intelligence, social

emotional learning, or sense of mastery, concepts that I haven't even mentioned today. These are things that I believe are highly valuable and that both we in the measurement and research community and our partners in schools and districts should be more mindful of. The test score accountability movement has pushed aside made many of these so-called "non-cognitive" or "soft" skills and they belong back on the front burner.

Let me talk about grit. It means pretty much what you think it does: "perseverance and passion for long-term goals."[12] According to Angela Duckworth and her colleagues "Grit entails working strenuously toward challenges, maintaining effort and interest over years despite failure, adversity, and plateaus in progress. The gritty individual approaches achievement as a marathon; his or her advantage is stamina. Whereas disappointment or boredom signals to others that it is time to change trajectory and cut losses, the gritty individual stays the course." Grit is measured very reliably with questions like, "I am diligent," "Setbacks don't discourage me" and "I have overcome setbacks to conquer an important challenge." Grit is related to, but not the same as self-control. Grit is **not** related to IQ. Duckworth and her colleagues have conducted a series of fascinating studies about grit and consistently find that it predicts success over and above other attributes like intelligence, self control, and grade

[12] Duckworth, A.L., Peterson, C., Matthews, M.D. & Kelly, D. R. 2007.Grit: Perseverance and passion for long-term goals. Journal of Personality and Social Psychology. 92(6):1087-1101.

point average. For example, grit predicted which freshmen cadets made it through summer training and their first year at West Point better than the other predictors. Grit predicted which children succeeded in the National Spelling Bee contest.

Now let me talk about a related concept: the quest for mastery. Over 25 years ago, my dissertation advisor, Benjamin Bloom, and several of his graduate students at the University of Chicago, studied highly successful adult swimmers, pianists, tennis players, neurologists and mathematicians, by conducting intensive interviews with these individuals, their families and their coaches and teachers.[13] They asked how the top people in their fields reached such high levels of success. One of the most notable findings from this vast study, described in a book called *Developing Talent in Young People,* was the huge investment and commitment of time, energy and effort that the experts and their families made in developing their skills. But it wasn't just hard work that led to success: it was focused, deliberate and strategic effort. A recent book, by Geoff Colvin, *Talent is Overrated*[14], makes these same points, arguing that success depends not solely on either talent or luck, or some combination of them, but instead on deliberate practice. Many hours of practice are necessary to reach a high level of success (some suggest the "ten year rule" or 15,000 hours), but reaching elite status

[13] Bloom, B.S. (Ed). 1985. *Developing talent in young people.* New York: Ballantine Books.
[14] Colvin, G. 2008. *Talent is overrated: What really separates world-class performers from everybody else.* New York: Portfolio

requires "practiced designed specifically to improve performance," often with the help of a teacher and with continuous feedback.[15]

About 20 years ago, another University of Chicago researcher, Mihaly Csikszentmihalyi wrote a wonderful book called *Flow: The Psychology of Optimal Experience.*[16] As you probably know, flow is that state of mind when you are fully immersed, focused, energized, engaged and involved and feel somewhat outside of yourself. Back when I read the book, I was a long distance runner and had experienced that amazing sensation of flow, "in the zone." Runners, rock climbers, musicians, computer programmers, writers, artists and yes, researchers and psychometricians, can and do experience flow. I really liked the fact that a respected academic studied this phenomenon, gave it a name, and built a theory and program of research around it. In the many years since I read the book my thoughts have returned to this concept again and again. I think about how important the concept is to me personally, I think how important and potentially useful it is in making the workplace both more enjoyable and more productive, and I think how crucial it is for children to have these sorts of optimal experiences that can encourage, motivate and sustain them. These experiences can change

[15] Colvin, page 66
[16] Csikszentmihalyi, M. 1990. *Flow: The psychology of optimal experience.* New York: Harper and Row, Publishers.

lives. As Csikszentmihalyi said "It is the full involvement of flow, rather than happiness, that makes for excellence in life."[17]

So here are three psychological constructs: grit, striving for mastery and flow. They are clearly interconnected. In his popular book, *Drive: the surprising truth about what motivates us*, Dan Pink argues that the experience of flow will motivate us to strive for mastery and become grittier.[18] These constructs are measurable. I think that they are also teachable, or as we say at IES, malleable. They are valuable in their own rights and they are also very likely to mediate success in school.

I've been trying to argue that the testing and measurement community has much to add to school improvement efforts. Good measurement, whether it's of achievement outcomes, important non-cognitive skills, or of school and classroom processes that lead to improvement has a big role to play in these efforts. So I am asking you to engage with practitioners and policy makers: build better large-scale assessments, help make teacher evaluation systems more useful, re-imagine NAEP for the future, and bring the so-called non cognitive skills back to the front burner. Use your technical skills to measure what's important, communicate across stakeholder groups, and advance school improvement.

Thank you.

---

[17] Ccikszentmihalyi, M. 1997. *Finding flow.* New York: Basic Books.
[18] Pink. Daniel H. 2009. *Drive: The surprising truth about what motivates us.* New York: Riverhead Book.