The Practice of Education Evaluation Today.

John Q. Easton. American Evaluation Association, October 18, 2013.

Washington, D.C.

Good afternoon. Thanks Jody and Tom for the invitation to speak today and for your patience with my scheduling problems.

The title in the program for my talk is "the practice of education evaluation today: a federal perspective." I need to tell you right off that I am not sufficiently knowledgeable to address such a broad topic. I am deficient when it comes to state-of-the-art evaluation theory and practice and am fully aware that you know more about this than I do. I am also not going to represent research and evaluation policy across the federal government. What I want to do is describe how I am thinking and explain how we're beginning to change inside the Institute of Education Sciences.

Part of the Department of Education, we are the federal government's home for education research, evaluation, statistics and assessment. IES was established in 2002 by the Education Science Reform Act. Perhaps our best known program is the National Assessment of Educational Progress, NAEP. We also collect and report education statistics, we conduct longitudinal surveys, such as the Early Childhood Longitudinal Survey and the High School Longitudinal Survey. This work is conducted by the National Center for Education Statistics, NCES.

The work that is most relevant today is conducted in the other three centers: two research centers, The National Center for Education Research and The National Center for Special Education Research, and the evaluation center, called the National Center for Education Evaluation and Regional Assistance. The two research centers make grants to education researchers. The evaluation center, among many other activities, conducts rigorous evaluations of federal and other education programs. Many of the evaluations are impact evaluations and many but not all of these are randomized control trials, using sophisticated clustering and blocking. The research centers support many types of studies, but an implicit goal has always been to move research along a "pipeline," from development to efficacy and effectiveness testing, emphasizing the use of RCTs. The idea of "establishing causal linkages" is in IES's DNA.

Right after Labor Day this year, Gina Kolata, the eminent science reporter for the New York Times, published a story called "Guesses and hype give way to data" in a special issue of the Tuesday Science Times called "Learning What Works".[1] Kolata extolled the use of clinical trials in education, suggesting that we are finally "catching up" with fields like medicine, health care and the development of new drugs to treat and cure diseases. IES clearly has played a big role in promoting these rigorous studies through our grant competitions and our evaluation contracting.

[1] http://www.nytimes.com/2013/09/03/science/applying-new-rigor-in-studying-education.html?pagewanted=all

In the four and a half years that I've been at IES, I've read, listened to and thought a lot about the RCT as the "gold standard" for education research and evaluation and about the emphasis that IES has traditionally placed on this method. I want to describe some of that thinking today.

First, let me tell you a little bit of my background. I came to Washington from Chicago where I had worked for more than 30 years with the Chicago Public Schools. From 1997 to 2009 I worked at the Consortium on Chicago School Research at the University of Chicago in close partnership with the school district. We worked hard to be partners with the district and hoped that our work would help guide school improvement. We strove to remain objective and independent, yet we still wanted our efforts to help the district as it charted its course. We conducted some studies at the district's request but most resulted from a periodic research agenda set to probe major issues and to build a coherent knowledge base about school improvement in Chicago. Our target audience was the school district, the civic community and the broad public. Through these experiences in Chicago, I came to believe that partnerships between researchers and practitioners are a powerful lever for making research and evaluation relevant and usable and as a result, IES has begun several new grant programs to develop these partnerships.

Most of our research in Chicago was highly descriptive and focused on naturally occurring variation across schools and communities. We did not do experiments: we

provided detailed descriptions of what was happening in schools across the city, displayed the variation and showed how that variation related to other factors.

Some of our research in school improvement is described in a book called *Organizing Schools for Improvement: Lessons from Chicago*.[2] The book lays out our evidence that five key organizational factors drove sustained school improvement in Chicago elementary schools during the 1990's. School leadership is the first of these and this leadership provides instructional guidance, and promotes professional capacity of staff members, family and community involvement and the school learning climate.

When I got to IES, I became quickly aware of a different paradigm. The thinking goes like this: schools improve by adopting proven practices and by faithfully implementing effective programs and interventions. IES has largely been defined by this approach: develop an intervention, program or tool, test it rigorously, and then scale it up. If it doesn't scale, poor implementation is at fault.

Charles Payne, a friend and former colleague of mine in Chicago, wrote a book several years ago called *So Much Reform, So Little Change.*[3] Charles described low performing, very poor schools in Chicago, ones that we later called "truly disadvantaged." Following a 1989 decentralization of authority in Chicago these schools gained considerable financial resources to fund school improvement efforts. They were

[2] Bryk, A.S., Sebring, P.B., Allensworth, E., Luppescu, S. & Easton, J.Q. 2010. *Organizing schools from improvement: Lessons from Chicago.* Chicago: University of Chicago Press.
[3] Payne, C.M. 2008. *So much reform, so little change: The persistence of failure in urban schools.* Cambridge, MA: Harvard Education Press.

besieged by sales representatives trying to sell their "proven" practices, materials, and curricula. Few of these interventions succeeded or even took hold. The schools lacked the basic human capital resources to implement, to monitor, or to ensure coherence or consistency across disparate programs.

In a paper called *Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Evaluations*[4] Dick Murnane and Richard Nelson argue that RCTs are great for validating interventions that can address and ameliorate very specific problems in schools. But low performing schools don't become high performing schools by implementing proven interventions. They become great schools by becoming learning organizations that chose carefully, monitor, discuss, analyze, adapt and refine. I am trying to broaden the IES view on school and system improvement.

I'm very sympathetic to today's emphasis on evidence-based policy and grant making, but I think that we need to be careful that our view of what constitutes evidence is not too narrow. It's hard to argue against using evidence in decision making, but we must recognize the value of evidence generated from studies besides RCTs. Just as we need a broader view of the school improvement process, I think we need a corresponding broader view of evidence.

[4] Murnane, R.J., & Nelson, R.R. 2007. Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Evaluations. *Economics of Innovation & New Technology, 16*(5), 307-322.

I am going to talk about two recent studies. One of these was conducted by my former colleagues in Chicago at the Consortium and the second was sponsored by IES and conducted by Mathematica. Before I start, let me assure you that I know I am comparing apples to oranges – first a formative evaluation or implementation study and the second an impact or summative evaluation.

In September, the Consortium released a study called "Teacher Evaluation in Practice: Implementing Chicago's REACH Students."[5] REACH Students is Chicago's teacher evaluation system, rolled out in the 2012-13 school year. The Consortium was funded by a local foundation to conduct this formative evaluation in "real time" by getting findings out quickly. This report describes teachers' and administrators' perceptions of the evaluation system. Data were collected largely through large scale surveys and a relatively limited number of interviews. The researchers knew this topic, having studied a pilot implementation of use of the Danielson teaching framework in classroom observations a few years ago in a similar quick turnaround mode.[6]

I think that this little study is powerful because it is so useful. It found that "overwhelming majorities of teachers and administrators believe the observation process supports teacher growth, identifies areas of strength and weakness, and has

[5] Sporte, S.E., Stevens, W.D., Healey, K., Jiang, J. & Hart. H. 2013. Teacher evaluation in practice: Implementing Chicago's REACH students. Chicago: Consortium on Chicago School Research. http://ccsr.uchicago.edu/publications/teacher-evaluation-practice-implementing-chicagos-reach-students

[6] http://ccsr.uchicago.edu/publications/rethinking-teacher-evaluation-chicago-lessons-learned-classroom-observations-principal

improved the quality of professional conversations between them."[7] It showed that teachers do not understand how measures of student growth are used in calculating overall ratings. It pointed to shortfalls in lines of communication within the district between the central office, building administrators and teachers. It quantified the amount of time required to conduct observations, pre- and post-observation meetings, and enter and manage data.[8] It described the tensions that teachers and administrators face when the role of evaluator and coach and helper is merged.[9]

This report provides powerful and useful feedback that will enable the district to rethink some parts of the system, fine tune others, and ramp up more. The Consortium is now analyzing more of the data they collected and examining how ratings, performance tasks and value add relate to each other. Because of the prior work in the pilot study, and levels of trust, the Chicago researchers have the background knowledge and capacity to do this work quickly and infuse their findings into the ongoing implementation efforts.

Here's a very different study. In early September, IES released a major study called "The effectiveness of secondary math teachers from Teach for America and the

[7] P. 2, Sporte et al
[8] The typical high school administrator spent 180 hours on these tasks last year.
[9] The study received a great deal of media attention, including a New York Times Sunday column by Brent Staples. http://www.nytimes.com/2013/09/29/opinion/sunday/principal-and-teacher-a-complex-duet.html?ref=teachersandschoolemployees&_r=0

Teaching Fellows Programs."[10] This is a well-designed and well-executed study about a controversial topic: non-traditionally prepared teachers – from Teacher for America and The New Teacher Project Teaching Fellows Program.

There are two separate studies – one comparing TFA teachers to traditionally prepared or other alternative route teachers; and the second similarly comparing Teaching Fellows to traditionally prepared or alternative route teachers. Both studies targeted math teachers in grades 6 to 12. The outcome was student value-added scores on the state math test for grades 6 to 8, and end-of-year assessments for grades 9 to 12.

This was a randomized control trial where students in the same grade and school were randomly assigned to either a TFA or Teaching Fellows teacher or a non-TFA/TF teacher. The results are pretty startling. Students of TFA teachers outperformed students of comparison teachers by about 2.6 months of math learning, even when the comparison teachers had more teaching experience. Overall, students of Teaching Fellows teachers did at least as well as comparison teachers, and when compared only to alternatively prepared teachers from non-selective routes, their students performed significantly better.

The study measured a number of factors that might be related to student achievement growth. Most didn't matter: selectivity of undergraduate college, number

[10] Clark, Melissa A., Hanley S. Chiang, Tim Silva, Sheena McConnell, Kathy Sonnenfeld, Anastasia Erbe, and Michael Puma. (2013). *The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows Programs: Executive Summary* (NCEE 2013-4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

of math courses taken, whether they majored in math or not. Only three did matter: experience, the number of hours that teachers spend outside of school taking course work, and teachers' knowledge of math content. In spite of the fact that TFA teachers had less experience and were taking more course work in the evenings and weekends, their students still outperformed the students of other teachers, both from traditional routes and from less selective alternative routes.

Here we have two very different studies: One is quick and descriptive and the other is a multi-year and it provides a trustworthy causal estimate. The first one provides a lot of information about what's going on in the teacher evaluation process but doesn't tell us if it makes a difference. The second tells us that TFA and TF teachers do at least as well if not better than traditionally prepared teachers. But it doesn't tell us why or what makes them better teachers.

I wonder if we can't do a better job of blending the two research- and evaluation approaches that I just described in a way that makes us more directly engaged in school improvement efforts. Can we develop models that provide useful feedback and trustworthy impact estimates? Can we learn not only if it works but why, where and how? Can we participate more closely in school and system improvement and demonstrate that we can be rigorous, relevant and useful all at the same time?

In my years at IES, I have come to value the importance of RCTs more than I did previously. The TFA/TF study served a very important function by definitively laying out

a set of facts that have been controversial and hotly debated. Yet, the study leaves a gaping hole: we don't know why. What is there about the selection, preparation and induction of these teachers that makes them do well? How can we replicate those conditions for other prospective or current teachers?

RCTs can give us confidence in making casual claims, but they are neither the sole nor always the best way to inform and create improvements in schools and in the education system. I think that there are two ways to approach expanding our research tools and methods. First, the RCTs themselves can be improved by providing more information about mediators and mechanisms – the "essential ingredients" of programs or interventions being studied. They can also do a better job of exploring the variations in outcomes that we almost always see in study results. Whether it's a dropout prevention program or a new drug, we need to know what about it makes it successful and we need to know for whom it works best, where and under what conditions. We need to design our studies to give more information about these questions. These topics are getting increasing attention from leading methodologists and researchers so I think we will make some progress on this front.

Beyond improved RCTs, I think that we also need to look elsewhere to help our work have greater impact, like the careful implementation study I described. In a short paper called *Broader Evidence for Bigger Impact,* Lisbeth Schorr lays out an argument

similar to what I am about to make.[11] We need to remain attentive to our causal impact evidence but at the same time draw more broadly from other research and evaluation methods, as well as from practice and experience. We use these sources of evidence to help us choose the right approaches to improvement, to adapt them with integrity, to continue to demonstrate that these efforts are leading to improved outcomes.

Like Schorr, I turn to the fields of improvement science and continuous improvement models that apply improvement science for guidance. Much of this work began in health care through organizations like the Institute for Healthcare Improvement under the leadership of Don Berwick.[12] In education today, the most prominent examples of continuous improvement come from the Carnegie Foundation for the Advancement of Teaching under the leadership of Tony Bryk.[13]

With major improvement efforts underway in teaching mathematics at the community college level, through programs called Statway and Quantway, Bryk and his colleagues are building a new research and development infrastructure. The improvement work is based on these six principles: 1. The work is problem-specific and user-centered; 2. It sees variation in performance as its core problem; 3. It emphasizes systems embedded in context; 4. It measures processes and outcomes; 5. Improvement

[11] Schorr, L.B. 2012. Broader evidence for bigger impact. Stanford Social Innovation Review.
[12] Berwick, D.M 2008. The science of improvement. JAMA*, 299*, 1182-1184*.*
[13] http://www.carnegiefoundation.org/

is based on disciplined inquiry, including RCTs, and 6. It accelerates improvement through networked communities.

I want to briefly mention some other interesting work that isn't explicitly improvement research, but could be cast in that framework. Stephanie Jones and Suzanne Bouffard from the Harvard Graduate School of Education wrote a great paper in an Society for Research in Child Development Social Policy Report called "Social and emotional learning in schools: from programs to strategies."[14] They argue that many schools need a new approach to promoting social emotional learning that is different from the "off the shelf" programmatic intervention model approach. Jones and Bouffard argue that social and emotional learning approaches must be embedded in daily practices and routines of schools and teachers. To quote them:

> We believe that schools need a continuum of approaches that range from routines and structures school staff and students use on a daily basis, to schoolwide efforts to promote respectful and supportive cultures and positive climates, to universal SEL programming for all students, to intensive services for students in need of the most support. Some schools' needs will demand, and their contexts will allow, that they utilize approaches from across the continuum, from everyday strategies to intensive interventions. Other schools may begin with the everyday strategies and add other components as the need and opportunities arise.[15]

[14] Jones, S.M. & Bouffard. 2012. Social and emotional learning in schools: From programs to strategies. SRCD Social Policy Report, volume 26, number 4.
[15] Jones and Bouffard, 2012, Page 12

This kind of approach to school improvement needs an analogous research and evaluation strategy to help guide iterative development, refinement and testing, what I would like to call research and evaluation for improvement. This could be similar to the improvement work that Bryk and his colleagues are undertaking or it could be different. In any case, it should be research that is distinctly conducted with improvement of practice as its goal.

I think that we need to build robust approaches to research and evaluation that are specifically designed to aid improvement efforts. These will need to flexibly combine various approaches, including formative evaluation strategies and RCTs and will need to be conducted collaboratively by practitioners and researchers.

At IES, we are calling for more relevant and useable research, for more collaborative partnerships with practitioners, and for more research and evaluation for improvement. In recent reports and hearings, the GAO and the Congress have told us that rigor isn't sufficient.[16] We need to continue pushing for more problem focused solution oriented useful research and evaluation. We need your help in tackling today's pressing education issues. I mentioned two of them earlier: teacher evaluation and teacher preparation, but I could add school safety, the teaching of social, emotional and psychological skills, assessing student learning, pushing more students to higher levels of problem solving and deeper comprehension. Members of the education research

[16] http://edworkforce.house.gov/calendar/eventsingle.aspx?EventID=348064

and evaluation community have formidable technical skills. Let's use them to create new models so that we can be active participants in the school improvement process and help solve some of these pressing education problems.