MR. MARTINEZ: Robert F. Boruch is University Trustee Chair Professor of Education and Professor of Statistics at the Wharton School at the University of Pennsylvania with previous appointments at Northwestern University and the University of Chicago, an engineer by training. I found out last night.

Dr. Boruch's areas of research include the design of controlled field experiments, ethics in data access and surveys, program evaluation, survey research methods, and federal evaluation policy.
He serves as consultant to numerous government agencies and foundations. Professor Boruch.

MR. BORUCH: Thank you, Mike. There are two handouts. One, a paper published in New Directions and entitled the “Flight of Error,” contains relevant references which some of you may find interesting. The other handout is a hard copy set of overheads based on the PowerPoint slides that outline my remarks.

Much of what I'm going to say is actually icing on Jimmy's cake. But I will cover activity in the kitchen and its environs more generally. That is, I'll try to write some of the lessons larger. By way of sequel to Jimmy’s fine presentation, let me make a prequel pitch to the National Science Foundation on how to spend your money.

Years ago, the Rockefeller Foundation mounted controlled randomized field tests in the employment and training arena for minority female single parents, poor ladies, lots of kids, no education, and so on. In developing those trials, in mounting them, we discovered that not much was understood about negotiating MOUs, and memorandum of understanding, in negotiating solutions to some of the problems, the difficulties, in mounting trials, getting over the obstacles, developing the partnerships, as Jimmy mentioned, with service providers, community-based organizations, and the like.

Rockefeller put money into videotaping researchers talking to senior people in community-based organizations on the topic of randomization
in their communities, the why, how, obstacles, and the objections, e.g.
randomization is illegal, it's immoral, and it's not much fun. These were
different at the time from the ones that you ran into. I suggested that
Rockefeller invest in videotaping in part because the kind of presentation that
Jimmy made, and certainly his experience in negotiations, cannot be captured
well in print. I've tried to do so myself as well as in partnership in publishing
edited volumes with other people.

There is a reality to seeing people interact on the topic of design
and execution of trials that helps build capacity. The Veterans Administration
has also engaged in videotaping Q and A, discussion of scientific and values
issues in situ, in the context of AIDS prevention trials, just as Rockefeller did
years ago. It's probably worth thinking about, for example, William T. Grant
Foundation as a potential source of money to lay something out that gets
beyond the print media and gets beyond the excellent but temporary benefit of
a lecture like Jimmy’s.

Okay. Let's see how to control the slide display from this post. I
am an ex-engineer, and the thumb is on the left-hand side. I should know how
to do this. Yeah, I got it. All right.

[slide2] This presentation’s theme, the Flight of Error, refers to the
political origins of some of our interest in evidence as opposed to rhetoric.
Michael mentioned it in his opening remarks. Jimmy mentioned the political
aspects of it.
It's great fun trying to trace the origins of the idea of fair comparisons and experimentation, not only in this country but other countries. Our being assembled here, in the great state of Virginia, seemed an appropriate occasion to open up with a quote from Thomas Jefferson. The context is his notes on Virginia and his arguments against the state's making a law that would imply a preference for one religion over another.

Jefferson's essay from his Notes on the State of Virginia is long and turgid, as a matter of fact. One of the few eloquent lines in it turns out to be his analogy in talking about government enforcement or government enactment of a law for religion just as enactment of a law of gravity. That is to say, that law "of gravity is more firmly established than it would be had the government stepped in and made it an article of faith." Further, said Jefferson, in referring to Galileo's science: "Reason and experiment were indulged and error has fled before them." That is almost poetic in its character.

Laying out the big scientific questions can help to put these experiments into context. The National Academy of Sciences Committee on Scientific Aspects of Education Research issued in 2002 and the book, Social Experiments (1997) outlined them.

I'm using a questionary framework and will depend on work from the 1960s, and the middles ages, in what follows. Some of you who know the history of language will recognize that the word "questionary" refers to the person who, during the Spanish Inquisition, questioned the tortuee. Those of you who have been subject to OMB's Part may sympathize with this construal--
Laughter.

MR. BORUCH: --and to other demands that are regulatory in character for more scientific evidence or better scientific evidence in this context might sympathize with that reference.

There are half a dozen big questions here. The questions addressed by experiments being one of them, critical, but only one. I'll try to step through each one because failing to recognize the importance of the other questions may lead to us inadvertently shooting ourselves in the foot by overemphasizing experiments at the expense of some of the other questions. So this just puts things into context.

[slide 6] The first question is what is the nature and severity of the problem that we're interested in addressing, based on what evidence? For instance, in what senses, based on what evidence, are kids not learning math or science? The second question pertains to factors associated or correlated with the phenomenon and the interventions that influence the phenomenon: How indeed do we know or observe program compliance or deployment, level of implementation, fidelity and so on? The third big question is what works question or what works better, and how do we know? More about this anon, with special emphasis on randomized controlled trials. Finally consider cost effectiveness ratios which cropped up earlier on account of Larry Sutor's question, and the related issues of replication and accumulation.

If you look at the budgets for some of the federal agencies which support experimentation at times, you find that most of the budgets are heavily
oriented toward questions other than the third question, on effect. Now this is partly because you have to answer other questions before you get to the effect question. For instance, we should understand the nature and severity of the problem before we start inventing programs to resolve the problem. Similarly, one should understand what the program is supposed to look like and the extent to which it is in place, and not merely assume it's there. So my reading of the federal investments, some of the Foundation investments certainly, is that efforts to address the question about impact are always going to be in the minority.

[slide 8] As I said, answers to later questions depend on evidentiary questions to the earlier ones. For instance, having worked on the What Works Clearinghouse for five years to understand what the problems are in studies generated by other folks, including some NSF supported studies, we did encounter a fair number of problems with respect to inappropriateness or insensitivity or unreliability of the measures, measures of the outcome, i.e. quality in measuring the nature of the problem.

We also encountered ambiguities in the nature of the control condition and lapses in attempts to characterize it well. This problem is relevant to another one of Larry's points earlier. That is, lack of information on the control program, lack of understanding of what goes on in the control group, means that you may be running a pointless experiment.

Some of the recent work in the Head Start arena, for example, suggests that the effect sizes are actually pretty small because there are a fair
number of programs out there that have ingredients that are much like Head Start's. Consequently, you're comparing one apple to another apple as opposed to another kind of fruit. So the effect size is bound to be small if there are no differences between the intervention and the control.

[slide 9] There are a number of other related frameworks, other than the questionary approach outlined, that earlier one can use to organize one's thinking about these larger questions and fund research on other relevant topics. IES's Early Childhood program, for instance, has five goals. One of them is directed toward funding research on measurement and assessment. Goals three and four are related to what we're talking about today. They concern relatively pure tests, randomized controlled trials to estimate the effective programs under ideal conditions, so-called efficacy trials, and randomized tests of curriculum in more realistic field conditions, so called effectiveness trials.

FDA has its own way of looking at the world of science policy and development of course, e.g. Phase 1, 2, 3 (trials) and 4 (postmarketing surveillance) studies. I haven't been able to find anything, anything in NSF's documentation on science policy to help arrange one's thinking about what NSF does, and I suppose this raises a question for NSF folks. Is there framework like these yet at NSF? Or is its development underway? Developing such frameworks is a challenge for the institution, I think.

[slide 10] The aforementioned frameworks for structuring the way we think about funding research and development attend to evidence. But they do not refer directly to the temporal character of its production. For funders, as well
as researchers, "when" we know is as important as "how" we know what works, for instance. ONR and DARPA, at least, are extremely clever about giving 20 year time horizons for their work, i.e., being persuasive about the fact that it's going to take 20 years to properly produce results and understand the import of various results. Our colleague Finbar Sloane, AKA Barry Slaone, takes a big view and I look forward to his presentation later.

Now, that time horizon is okay so long as they can provide a few pearls, winners of some sort, early on and along the way. Understanding how to arrange a research portfolio so you don't have to wait the five years while you're trying to get tenure especially and failing to publish is as a big challenge. Understanding how to do this right is not clear.

[slide 11] Let me now return to the questionary framework and consider each item in the framework seriatim. In regard to the nature and severity of the problem, a big issue encountered in What Works Clearinghouse work, at least, is largely the measurement side. In particular, we can measure kids in the middle of the distribution pretty well, the average kids, but in the lower tail, kids who do really poorly, or on the upper tail, kids who do really well, it's tougher to develop standardized measures that are easily adapted to the target populations and in a way that helps index within year improvements in response to interventions. This is one of the reasons why it doesn't come as a surprise that the effect size, say even for the Iowa Test of Basic Skills, comes out, quote, "smaller than you might want it to be," or that the variance estimates are too large and you don't find statistical significance in comparisons.
The What Works Clearinghouse made a decision to rely heavily on standardized tests, nationally normed or at least regionally normed and the like, as opposed to locally developed tests that were tailored to suit the particular intervention that people were trying to assess in a randomized trial.

Outcome variables that are highly aligned with the intervention were not regarded as fair. This has caused some controversy. That is, we do not yet have generally agreed upon standards for judging “alignment” between interventions and the outcome variables that are supposed to reflect the intervention’s effect, nor standards on the proper relation between locally tailored tests (highly aligned) and the standardized tests that must usually be employed to judge the failure of the interventions in a larger context. Researchers such as Fantuzzo, McDermott, Porter, among others are doing much needed work on the topic but resolution is far from clear as yet. On top of this, WWC and others usually had to assume in a that the measurement in each arm of the trial was equally good; published reports usually don’t give test reliabilities for both arms of the trial.

A final measurement issue encountered in WWC involves the multiplicity of the outcome variables that people measure. In the medical arena simple variables such as being dead and not dead (and hazard rates, etc) are fundamental. Right? Simple to count. And maybe functional mobility comes next as the second class of Y variable. The education arena varies considerably in what outcome variables are regarded as important in various
grades and subgroups and for various education topic areas; we have lots of Ys out there.

Understanding how to test hypotheses with a multiplicity of Ys in ways that are powerful and do not just rely on Bonferroni confidence intervals or tests as dividing the alpha level by eight if there are eight outcome variables and then declaring significance based on this very stringent criteria is not the right way to go about business.

[slide 13] But there are new developments and there are some references in the paper to that I think. Manipulatable factors, you already had a nice illustration based on that early work by Barbara Heyns, which is remarkable. Thank you for that reminder. It was lovely to be reminded about. And you all have the general idea about that.

[slide 14] What programs are deployed? Major hole. In most of the experiments and quasi-experiments reviewed under the auspices of the What Works Clearinghouse, I think there are about 80 interventions plus, which have been examined, each of which were examined based on either randomized control trial evidence or quasi-experiments.

[slide 15] The big hole in the literature here is sort of level of implementation, fidelity and the like. There's nothing remotely resembling sort of IRT technology for the Y variable for that I and C variable, in other words, the deployment, the implementation, the dose level. One often gets a list of ingredients and you might get hours spent on each ingredient at best, but
apart from that, there's actually very little out there that's coherent and which makes for major difficulties in replication.

Typically what happens in the control group is often in the peer review journals left out entirely. Somebody says the phrase "business as usual" in one way or another usually occurs, but one often doesn't know what the new intervention is being compared to.

Oh, let me mention this guaranteed failure and guaranteed success in a trial. Certainly, you can guarantee success by aligning the outcome variable very tightly to the intervention. Teaching the kid how to read Sanskrit and then testing in Sanskrit or teaching the kid about that paragraph in Thomas Jefferson and then asking questions about that paragraph in Thomas Jefferson as opposed to a control group which has never seen Sanskrit, never heard of Thomas Jefferson, and so on.

In the drug arena, so-called C trials now being mounted by drug companies, in which there are lots of small trials which are then combined statistically to make declarations about whether the drug works. Typically those C trials compare the new drug to what amounts to an obsolete drug. They find physicians who are actually out of date using older drug and ask them to participate in the trial, and they're comparing this new drug which is say second or third generation to a drug that's at the zero generation being used by the physician.
It's another way of guaranteeing success, at least for some of the trials run by the pharmaceutical companies. And certainly there are lots of ways to guarantee failure which I'll get into a little bit later.

[slide 16] On the What Works, I've already said the answers to this question are conditional on answering the earlier ones. The big ticket--the thing that's been a surprise to me over the last five years is the explosion in growth of these cluster randomized trials, group randomized trials, place randomized trials of the sort which I'll talk about a little bit more.

[slide 17] And let me, in part because some of these issues are actually quite big. Understanding what the World Bank, for example, thinks about controlled experiments in education is important, at least for Bank people and for the people who get loans from the Bank to concoct and employ programs that are supposed to prevent dropouts or enhance the literacy of kids or the scientific abilities.

In talking about the functions of trials in the context of evaluation a couple of years ago, two economists, one from the Bank, one from MIT, made presentations on the virtues of randomized allocation, the fact that guarantees at least up to a point for a comparison, statistically unbiased estimates of effect, none of the bustling and groping over three, four, five level models based on observational passive data, which cannot be explained to a teacher or parent, and so on.

Somebody in the audience, a Bank person--I was a commentator on the papers, both of which were pretty good--somebody in the audience stood up
and said, look, you economists--you can see where this is going--right--"you economists."

[Laughter.]

MR. BORUCH: --should behave more like astronomers and make better forecasts of what people would have done in the absence of the program in order to estimate the effect of the program on the kids or the parents or whoever instead of fooling around with people's lives and inconveniencing everyone by doing these randomized trials.

Just make better forecasts. Well, there was a kind of stunned silence--- deer in the headlights phenomenon. Nobody can argue with the aspiration to forecast better. But it's like asking someone to levitate. It's very hard to do. Right. Especially in very noisy environments.

Moreover, that ability is domain specific, and by that I mean years ago the--I think it was the Oklahoma City Police Department decided to test the intervention called "bullet proof cloth," Kevlar, to determine if that intervention was effective. Police departments had a natural interest in understanding whether the claims were accurate.

They got a supply of the stuff from DuPont, a woman who's--I regret I can't remember her name--who actually invented it, had major responsibility for its invention--supply of the stuff from DuPont. It was like heavy drapery at the time, and in their tests, they took a pig out to the
parking lot behind the police station, dropped the cloth over the pig, took out one of these high caliber 357s and fired at the pig.

Now, the intervention invariably worked in the sense that there was literally no bloodshed. All right. Main outcome variable of interest. Pig actually died of internal hemorrhaging, but that wasn’t on the outcome variable.

There was no bloodshed. There was no penetration. The intervention worked in that sense. Well, ask yourself the question, how many naked defenseless control group pigs do you think you need in order to persuade yourself that the causal inference that the intervention worked is accurate?

Well, the answer then, as now, is none provided you're willing to make some assumptions. Weapon fired properly. Weapon was pointed in the right direction, and, pertinent to the World Bank issue, the ballistic trajectory equations developed by Galileo were still operative in the parking lot behind the police station. Could be a context effect, but it's implausible.

Neither NSF nor the World Bank can wait the 400 years for those equations to be developed by people as good as Galileo. It may be possible in some context to be sure, some domains in education including science, technology and engineering education, but you have to concentrate and dedicate some real effort to looking for those areas where the predictability is so good, say over the short term, for certain kinds of groups, that we needn't bother or we needn't take--we needn't invest an awful lot of money in these control trials.
Okay. That's still a kind of science policy issue that's worth thinking seriously about.

Consider now the last big question in my list of big questions: what is the cost effectiveness ratio of alternatives that have been tested? I think either Stiglitz, Nobel Prize winner announced that economists do no cost benefit analyses of their cost benefit analyses. Despite this delicious self criticism, the idea of estimating costs of alternatives and relating these to effect size is attractive. Of course, it's easy to compute back of the envelope ratios. And indeed, the U.S. Government Accountability Office probably does about 500 of these things each year because it's required to do so by the Congress.

However, one of the difficulties in such work has been that economists have often not been able to rely on dependable evidence addressing those first three questions being addressed properly in order to generate defensible cost effectiveness ratios. If you don't do the experiment, you don't have a defensible estimate of the effects. If you don't know the magnitude of the problem or its cost, it is difficult to develop ratios beyond the sample employed in the narrow experiment at hand.

It's only recently that outfits like the Washington Policy Institute have been commissioned by Washington legislature to produce cost effectiveness analyses of the programs that are supported by the state of Washington. That state seems to be a real leader in the sense of providing results of conscientious review, and estimates of ratios based on randomized
results where possible, to legislative staff and legislators. The process is routine rather than episodic so that people have an opportunity to learn about the nature of such evidence and how to use it. Needless to say, however, we do not yet have a coherent body of evidence on when and under what conditions such evidence is used and in what sense. Younger colleagues could make distinctive contribution by generating such evidence, especially on the topic of when political and social values trump dependable evidence about negative ratios.

During my tour of duty with WWC, I encountered NO reports on trials reviewed by the What Works Clearinghouse that included thoughtful reports on cost effectiveness ratios except in the drop out sector. Peer-reviewed journal articles and unpublished reports that are the basis of WWC reviews, for instance, don't routinely include estimates of the cost of the interventions being compared in addition to numbing detail on the multivariate analysis of variance. There's not one dollar sign appearing anywhere in many academic reports because it's somehow regarded as, well, undignified or unscientific to do so.

Okay.

[slide 19]  Let me next cover the related topic of replication and cumulation of evidence based on randomized trials. The trialist aspires to get one trial done at least and many then try to replicate it in the interest of confirming or disconfirming initial results. But what indeed does replication mean? Brian Flay other colleagues have, are developing a tract on standards for deciding
when something is replicated for the Society for Prevention Research. Based on an earlier paper published in Prevention Science, the effort was undertaken because it is not always obvious what is a replication is, what's a partial replication, a dependable replication, a failure to replicate and so on.

There's lots of dimensionality embedded in the idea of replication. Replications are often sequential rather than simultaneous. Well, is that a replication or not? It differs with respect to time. Replication implies identity or similarity of samples and sampling procedure (eligibility for the trial), measures of the outcome, trial design including statistical power, and analysis method.

We use the word "replication" piously, but it's not always obvious when a replication is indeed that or what we mean by it. This has generated some squabbles in the crime and justice arena. Steve Levitt, the same Levitt I think who did the thing that you referred to, has a book entitled Freakonomics in which he's very clever, good writer, and has some interesting ideas about how to reanalyze existing data. Levitt said in print that he could not replicate some of Lott's analyses of the impact of "shall carry" weapons laws on homicide rates.

Lott maintains that the legal ability to carry weapons, concealed leads to a reduction in homicide rates.
Lott last year filed a suit against Steve Levitt for making that statement because he said it impugned Lott’s motives. One could construe the declaration of a failure to replicate as a suggestion that the analyst fabricated or falsified the data.

Now, I hope fervently that the definition of replication does not land in the courts because I think it would be a mistake. It’s the scientific community that ought to be defining these things. The fact that it has gotten to this stage, at least in the criminological research arena, suggests to me that we may see more of the same because the stakes are high.

For example, in the education arena, it’s not inconceivable to think about major global publishing companies that sell curriculum packages threatening to sue or merely threatening to tie you up with lawyers for the rest of your life because you failed to find an effect or failed to discover significant effect based on your replication of a trial that the publisher did.

If and when we replicate controlled trials in a scientifically acceptable way, how then do we understand results of the assembly of studies, i.e. cumulate and analyze the evidence. Cumulation and analysis are aims of outfits like the What Works Clearinghouse, the Campbell Collaboration, and Cochrane Collaboration. Similar efforts are multiplying out there, e.g. the Best Evidence Encyclopedia, Blueprints, NREPP, and so on. People are taking seriously the fact that evidence counts, quality of evidence counts, and different outfits are generating different ways of collecting the evidence and
sticking it together to reach conclusions about what works, and more often, what does not.

Let's see, the Cochrane Collaboration was created in the health care sector in 1993. It now has, oh, something like 5,000 systematic reviews of controlled clinical trials in the health care sector.

As I mentioned, the What Works Clearinghouse probably has about 80. Some of the early work on assembly of studies and analysis of the assembly was, incidentally, sponsored by National Science Foundation. This includes, for instance, support on the sub-technology called meta-analysis, e.g. putting the results of these different trials together to try to come up with an estimate of effect based on lots of small trials as opposed to one big one.

[slide 21] At the science policy level, the ability to cumulate and analyze assemblies of studies invites one to think about the odds on failure and success in the aggregate. If you ask yourself the question, as some NSF review panels used to do routinely, what's the probability that this thing is going to go, succeed? Or, what's the probability that it's going to fail? What indeed are the empirical odds on success? What do we understand about the odds? Should we think about that in the context of funding innovative human enterprise as in the medical science technology or engineering arena?

In thinking about odds on success of course, we need to recognize publication bias. You know lots of reports on interventions that fail go into a file drawer. You don't see them in the peer reviewed journals. The situation in
education research is likely to be at least as severe as in the other social sciences and medicine.

In medical efforts to handle this issue, for instance, there's a new journal entitled Journal of Negative Results in Biomedicine.

[Laughter.]

MR. BORUCH: This, a serious journal, was developed mainly because so many people shoved those negative results into a file drawer and failed to report them. This habit, in turn, led to careers being wasted, or long stretches of time being wasted by young people who in going up a blind alley discovered at the end that some of their older colleagues had gone down the same blind alley ten years earlier but had failed to report those results.

[slide 23] There's an enormous waste of human capital in this arena potentially. Partly for this reason, partly in the interest of informing science policy, some folks are tried to do a better job developing, arranging one's thinking about odds on success, failure, and in between. A rule of thumb here, dating from the '70s, is based on the medical literature, including work by Fred Mosteller and other folks. Roughly speaking, they found in examining controlled trials in the innovative surgery arena, that the new intervention worked in the sense of being a success, clearly superior, about 20 percent of the time. About 20 percent of the time, you had negative effects, killed more patients that way than you would another way. And the remaining 60 percent of the time didn't make much difference.
Now that, at the time, was not a bad basis for a rule of thumb for trying to reckon or anticipate in that arena what would happen. But we lack a substantial and coherent body of evidence and ways to arrange one's thinking about that evidence in this education research sector.

The notion of cumulating evidence is integral with the idea of cumulating methods that generate better evidence, of course. By way of giving credit where credit is due on this account, let me acknowledge the 1974 book that Henry Riecken, I, and others put together under the auspices of the Social Science Research Council. Entitled Social Experimentation (Academic Press), it included coverage of education experiments, including structured abstracts on those experiments. It was sponsored wholly by the National Science Foundation. There's a subsequent book entitled Experimental Testing of Public Policy published in 1975 to which NSF also kicked in money. NSF's investment in this arena actually was substantial early on. It is not well enough recognized now, I might add.

Of course, the early investments by NSF, and by the scholars who generated important work based on the support, are now being superseded by others. Steve Raudenbush, Tony Bryk, and their students and colleagues are important in this advancement. The recent work by people at Mathematica Policy Research, Manpower Demonstration Research Corporation, American Institutes for Research, and small firms such as Analytica, are no less important for their applying the ideas and discovering where the ideas are insufficient. Development of better approaches to assuring statistical power
in complex experiments has expanded and gotten a lot easier on account of colleagues’ efforts. Apart from people who I’ve mentioned, Howard Bloom at MDRC, Mark Lipsey at Vanderbilt, Larry Hedges and Spryros Kanstantopoulos are among the admirably productive in this arena.

Although there are early precedents, these scholars and organizations have contributed notably to contemporary experiments in which entire school districts or schools or classrooms are the units of random allocation. The growth in these place-randomized trials, this increased prevalence, was pretty strong beginning in the 1990s. There has been a really big jump since 2000, partly as a consequence of Grant Foundation, IES and others, including some NSF supported cluster randomized trials. The chart illustrates the growth. The special May 2005 issue of the Annals of the American Academy of Political and Social Sciences illustrates the scope of these beyond the education sector.

How much time do I have?

MR. MARTINEZ: 15 minutes, 20 minutes, something like that.

MR. BORUCH: Oh, okay.

MR. MARTINEZ: Plenty of time.

MR. BORUCH: Yeah, well, we have discussion, too; right?

MR. MARTINEZ: Following.
MR. BORUCH: Following, okay. Let me use the remaining time for a few observations on recent work in comparing the results of randomized trials against nonrandomized trials, the choppy pedigree of randomized trials, and on power of complex trials. The business of comparing outcomes of randomized trials to non-randomized trials for roughly the same intervention and roughly the same target population and so on has become a bit of a cottage industry among some of our methodological colleagues.

[slide 28] The activity nonetheless has early origins. If one looks carefully (not with Google), one can find a 1950s issue of the British Journal of Agricultural Research in which Yates described randomized control trials on spreading manure. The aim was to determine whether it enhanced the potato crop yield. This experiment had followed quasi-experimental work in which farmers' practices and crop yields were observed through passive surveys. The outcome—the Y variable here—is the crop yield. Basic regression analyses, ordinarily least-square stuff, was used to determine whether the regression coefficient for spreading manure was positive, negative or close to zero. It turned out to be the case that the non-randomized trial produced negative estimates. In other words, it made the program, the intervention look harmful. The randomized trial produced positive estimates which were later replicated. Okay.

This is one of the earliest examples I could find of this kind of empirical comparison being made mainly in the interest of doing a better job estimating the effect of spreading manure on the potato crop.
Certainly it's easy to show with simple model building using conventional methods, say analysis of covariance in passive observational data, that such methods under certain conditions and assumptions about the conditions can make programs look harmful. Campbell and I did this decades ago to demonstrate that it's possible to produce negative estimates of program effect, when, in fact, in fact merely useless. Under other assumptions and conditions, one may make the program appear to have positive effects when it is merely useless.

The bottom line for the quasi-experimental efforts is: You let me make the assumptions and I can estimate the socks off the effect of anything in the world. Imbedded in those model-building exercises are assumptions that actually produce the negative or positive estimates. Anyway, in the long run, it's not unreasonable to argue that comparisons of this sort are going to be a fundamental part of this science that we're in, partly because we can't always do the randomized trials. Because at times we have to rely on the non-randomized trials, we ought to understand something about their potential biases, what direction they go in, and potentially why.

So these empirical comparisons are important. Beyond that, getting more information about the nature of self-selection in non-randomized trials is absolutely essential. Otherwise, we will continue to commit the same mistakes that we've made in the past for years on the non-randomized trial side: failing to predict well what happens in the absence of the intervention,
and what would happen in the case of self selection in voluntary programs (this latter is Heckman’s point about limitations in trials).

[slide 29] Okay. Consider next the pedigrees and choppy history of randomized trials. You probably don't need this. By choppy history, I mean that the growth of these trials differs across time and domain. All right. There were actually very few really well controlled randomized trials in the medical sector before the Salk vaccine and Streptomycin trials in the '50s. The Salk trials incidentally included both quasi-experimental designs as well as a randomized experimental design. The experimental design involved second graders being randomly allocated with parental consent, to the placebo versus Salk.

In the quasi-experiments, run in about half the states, all second graders got the vaccine and first and third graders were used as a control condition. This was under the assumption that if you average a first grader with a third grader, you get a second grader. Right? This sounds terribly simple-minded, but this is the way it worked out. And indeed, results of both kinds of studies differ.

Iain Chalmers, Arild Bjorndal are looking into early history of randomized trials in education. In regard to STEM, the earliest trial I can find was mounted by Herb Walberg and colleagues on Harvard’s Project Physics in the late 1960s. Why this remarkable precedent failed to trigger further interest and investments in randomized trials in the STEM sector deserves attention from contemporary historians. Of course, the medical area eventually had an
FDA to assure fair comparisons. It was not until recently that education got an Institute for Education Sciences to provide strong influence if not regulatory control to assure fair comparisons. Beginning in roughly the year 2000, things really seem to have taken off on account of IES. At least 60 randomized trials have been mounted including some in the STEM arena.

Part of this acceleration in conducting trials is doubtless attributable to No Child Left Behind. Part of it is attributable other stars in the cosmos coming together. For instance, our brothers and sisters over in the Institute for Education Sciences are sponsoring experiments through both contracts and grants.

In the chart, the NCER represents the education research wing, and has its own commissioner, Lynn Okagaki, and the NCE is the educational evaluation wing whose commissioner is Phoebe Cottingham. Both wings support, provide, push money at randomized controlled trials of various kinds. Two-level trials are typically ones in which classrooms, for example, are randomly allocated to different regimens, different teaching methods, different curriculum packages. Kids within classroom are not randomly assigned but measures on them are taken and analyzed in the hierarchical context.

[slides 31 & 32] Three-level models may involve schools being randomly assigned to different regimens. Entire schools that are willing and eligible pick up a package such as Success for All, or do not pick it up based on the random assignment. Classrooms are nested within schools, and kids are nested within classrooms. There may or may not be random assignment at
those lower levels. The three-level multi-site trials might involve a blocking variable or a stratification variable like school district. We got district one, district two, district three, and then schools within districts are randomly assigned, and then you get the kids in the classrooms within the district.

[slide 33] As these study designs get more sophisticated, the technical burden of estimating statistical power becomes increasingly complicated. You stick layer on layer on layer and make the design more sophisticated in the same sense that engineers have to estimate an object’s velocity relative to different reference systems (relative to a roadway, the earth’s rotation on its axis, the earth’s rotation around the sun etc.). Jessica Spybrookka did a lovely dissertation at the University of Michigan in 2007 on power analyses in this complex environment, how to do them, and whether the power analyses done by the original people doing the proposal were on target.

It turns out to be the case that the academic community's proposals were often for under-powered trials, partly because the academic community didn't ask for as much money as some of the firms do, like Mathematica, American Institute for Research, and so on and so forth, who just know how to ask for more money for these larger-scale trials which assure that the effects, if they're there, are going to be more detectible.

[slide 35] On the capacity building side, for these sophisticated designs, we have lacked regularly accessible software to estimate statistical power as part of the experimental design. To his credit, Bob Granger, President of William T. Grant Foundation, came up with the idea that one way to develop increased
capacity in this arena is to support the generation of software, make it publicly accessible, and provide workshops for reasonably senior people to learn how to use it. This so as to give large-scale randomized trials a kick-start, but a sophisticated one. The foundation site still includes that software; it's relatively easy to use, explain, and comes with documentation. My young and very bright colleagues, such as Herb Turner, CEO of Analytica, know about this and use it, but also exploit other software (SAS for instance) so as to tailor the trials design to accommodate local constraints and limitations on resources.

Let's see. Jimmy talked about minimum detectible effect size, in effect, and also alluded to analysis. On statistical analyses, let me reiterate one or two lessons drawn from the Clearinghouse as well as from the W. T. Grant Foundation. I served, until last May also on the board of trustees for the Foundation so I got to look at a lot of proposals in this arena.

[slide 37] One of the biggest errors in some research areas, at least up until, say five years ago, had to do with failure to do intent-to-treat analyses. That is to say, you should analyze the data on units as you randomized them as the first level of analysis as opposed to analyzing data only on those who stayed with the program. So, in a randomized trial for example, even if the kid dropped out of the program or the teacher dropped out of the program or the intervention group, you still treat them as if they were in there, analyzing the data, recognizing that it's a weakened form of treatment because they're dropout. But the comparison between that experimental group and the control group are fair on account of the randomization process. And the statistical
mechanics of randomization on the technology is appropriate to estimate Type I errors, for example, when you treat the data that way (intent to treat) and not otherwise.

One of the biggest odd problems in looking at older studies in the What Worked Clearinghouse how to do with the fact that for some cluster randomized trials, the researcher would do something like randomly allocate school A to the intervention group and B to the control condition.

The researcher then got 200 kids in each school so the total sample size is 400. And the t ratios and the analysis of variance were all based on the kid as the individual unit of analysis. Now, that's wrong because if you rely on the randomization theory to test the hypothesis, you've only got--you've got zero degrees of freedom. This is because you had just two units that were randomly allocated. Reports on such studies were in peer reviewed journals and in institutional reports despite the early recognition of the error in books, by Leroy Wolins in the 1970s for instance, and others (see the Kempthorne aphorism in the Analysis chapter in my book, Randomized Experiments).

Correcting this and other kinds of error for dozens of studies turned out to be a technical challenge all by itself because you don't want to throw all of the data away. Sometimes the researchers had six, eight, ten schools, randomly allocated. But they just mis-analyzed the data and correcting for that mistake became part of the Clearinghouse's challenge. And a lot of
good documentation on this is given on the WWC Web site (Larr Hedges contributed notably to this).

[slide 38] Let's see. The ending slide, before the appendices, quotes a great pundit, Walter Lippmann. He was criticizing Roosevelt for, in Lippmann's opinion, being not an entirely honest experimentalist. The word "experimentation," as in social experimentation at the time was pretty popular. Don Campbell capitalized on it, as I do. The Lippmann declaration in may paraphrase is: Unless we are honestly experimental, we will leave the great questions of society to the ignorant opponents of change on the one hand and ill informed advocates of change on the other.

This is another one of those lines from the political arena that's almost melodic in its character. Okay, I have shot my bolt. A potentially embarrassing number of them

That's it except for all those lovely equations at the back of the slides. Doubtless, someone will be eagerly interested in figuring out what they mean.

MR. MARTINEZ: Thank you.

[Applause.]

MR. MARTINEZ: We are a few minutes ahead of schedule which is terrific. So we have nearly 20 minutes for questions if we want to take that long.
MR. SCHWEBACH: With all this potential error, is there any particular kind of study that we should be looking to that would have less error most likely? I mean where do you start if everything is complex?

MR. BORUCH: You're looking for good models, good examples. Some of those experiments in which the data had been routinely reanalyzed and looked over by other people including harsh critics are good place to start. Jimmy mentioned Tennessee STAR, the Tennessee experiment, reanalyzed by among others former Chief Economist at the Bureau of Labor Statistics Alan Krueger, who also writes for the Times, economist at Princeton, reanalyzed by Fred Mosteller, and I think Dick Light also. That's not a bad model.

They couldn't find any major flaws. If you take a look at the recent, most recent proposals, actually even the earlier ones, submitted to Institute for Education Sciences, for example, for some of the control trials in that arena, by and large, they're well designed. Execution is another matter.

Footnote: You recognize that you got to have good people in the field. Delivery of these things is crucial.

But from the design side, getting a hold of some of those proposals and the criteria used to evaluate them would probably be a good thing to do. Technically, it's public information, FOIA, and all the rest of this business. As a practical matter, they're tough to come by.

I think they are more accessible when some of the big firms are involved because they are used to generating redacted proposals; that is proposals which won but which exclude proprietary information. All right.
Which they then distribute to other people. It's the large and even medium-size firms are not unused to doing that.

In the academic arena, we are not used to doing that. I certainly am not used to developing a redacted proposal. I mean just take the money part off and Xerox it up, send it out, if I do that at all.

Still other people including in the educational research arena including some trialists refuse to provide you with a copy of the proposal although they get money from a publicly supported agency for doing the research.

All right. So there are different customs for different people. I regard that as potentially unethical, but it's a debatable issue. Nothing is--one other kind of way of answering your question is nothing is flawless. It's alarmingly easy to find a twist, a wrinkle, a potential flaw in all of this stuff, but understanding how to just raise the level of quality of evidence, raise the level of conversation about it, decide what's a trivial flaw as opposed to what's not is the big ticket item.

Uh-huh. Yeah?

MR. YIN: Hi, Bob. I would just like to follow up on that question and narrow it to math and science and ask you in all of your reviews, whether you can describe a good randomized trial study in math and science and just comment or speculate on why we don't find more of them relative to other subjects?
MR. BORUCH: Let's see. The first place I'd look for good trials is on that What Works Clearinghouse Web site—all right—because if they exist, it's going to be there.

The searches through the grant literature as well as peer-reviewed literature, exploiting the invisible college, the networks, being neurotically conscientious about examining those reports was basically the routine for five years, and probably will be over the next five for the next round.

So the first place I'd look for interesting illustrations are ones that, studies which are deemed acceptable and dependable for among others I think early, oh, middle school math, and I think elementary school math is also in there. I can't recall science or any of the sciences. But recall the Walberg Harvard Project Physics trial as a reminder that there was research and reports on it prior to the ascension of Google and IES.

Now you have to remember also, however, that the Clearinghouse focuses on packages that are commercially available typically. This was in the interest of assisting the users out there, the school district people and the like, who had to purchase packages from major publishers.

Work of the sort that Jimmy was doing on the books might not fall into that category because you're not a commercial outfit that tells people how to spread the books around, and you don't have a copyright on the material, the lesson plans and the like as yet, but if you formed your own company and did that, then you'd be one of the targets for the Clearinghouse.
It's the production rate may and coverage of more programs may increase, but for every intervention or study that's actually, intervention that's examined by the Clearinghouse with respect to evidence, there are at least eight other interventions in math, science and other areas, reading, for which there are commercially available packages and for which there is no evidence at all, no tests worthy of the name. Might be before or after, but that's about the limit, and the Clearinghouse doesn't rely on before/after studies.

Larry, were you going to say something?

MR. SUTER: Thank you. Actually, Bob, I was trying to figure out what comes out of this? What you're doing is counting randomized experiments and you spend a lot of your time in your slides talking more about power than about effect size, and you never use the word "cause."

In fact, I even look in your equations, and you avoided the discussion of cause there, and your conceptual framework, the RFB's questionary framework, are questions about what we know and what works.

So I'm trying to figure out what, what these data are for? Where is--you never talk about the need for an overall framework of knowledge or of education or of school practice or of medical practice.

MR. BORUCH: Uh-huh.

MR. SUTER: Should we--now, maybe you don't have to because you're talking--you don't even talk about cause. So we're just talking about data. But I'm trying to figure out what, if we want more randomized
experiments, what do we want them for? What questions are they really answering?

MR. BORUCH: That's--that third bullet--what works, what works better--is a fairly narrow question. Let me get to this cause thing first. As Michael mentioned, I'm an ex-engineer--all right. For me it's more natural to talk about fair comparisons, unbiased estimates, than it is to talk about cause. All right.

Furthermore, beyond the ex-engineer thing, I must say it's been less complicated, less difficult to explain these controlled trials and their purpose to people when I avoid using causal language than when I say, use the phrase "fair comparisons," people, judges, state legislators, staffers. The idea of fair comparison or fairness in the comparison is a little more transparent than educating whether one thing caused another.

So for me it's easier to sell the idea of control trials by avoiding that causal language. Now, that's a fairly narrow mission, however. When you talk about the larger sort of building the knowledge in a particular arena, that's actually one of the reasons I enumerated those half a dozen questions as opposed to just the one. Understanding where to stick these things in the generation of knowledge is important scientific task as far as I'm concerned.

One of NSF's own, Barry Sloane will talk about this anon and probably will be eloquent in doing so. Paul Holland and Don Rubin deserve high credit for mapping the causal language onto the statistical methods that we use
to generate evidence about “what works,” a phrase that is loaded with causal connotations.

That's one of the reasons I respect and actually support or vote in favor of support of work like Barbara Heyns did years ago. Who supported that; do you remember, Jimmy?

MR. KIM: She was part of the original Jencks team that did the reanalysis.

MR. BORUCH: Uh-huh. Okay.

MR. KIM: I think it was at--might have been at--

MR. SUTER: It was a staying effect study that came out--

MR. KIM: Yeah, that's right.

MR. SUTER: --the old department of Ed.

MR. BORUCH: Yeah, right. Okay. All right. Now, how you stick all this stuff together is frankly beyond me. I'm going to leave it up to my students who have more stamina, more brains, more active synapses than I do, or to some of you folks who are thinking harder about these much, much larger issues.

Yeah.

MR. HAMOS: The question I wanted to go back to the previous comments you were making about interventions and what I'm wondering is from
your perspective, what is the level of the, in which an intervention is powerful enough to do RCTs?

So, for example, you were talking about packages. That can be thought as a very broad intervention. It has many subcomponents to it beyond just doing, you know, one of the curriculum materials or not.

MR. BORUCH: Yes.

MR. HAMOS: So what scale is it should we do innovations, that should go down the road of an RCT?

MR. BORUCH: Yes. Put in other words, the question is What intervention is worth testing in a controlled trial, what size, what scope, and so on? That's an exquisitely interesting question, an easy answer to which I do not have. Certainly thinking about that question before getting into a trial is important.

To get back to Jimmy’s example, is Eight books enough or should it have been 16 books? Why eight and so on and so forth? Why not four? Now, I know of no one who's done really hard thinking coupled with some empirical data on that topic. At what level, say tipping point for us; right? At what point do you decide, oh, now it's worth, the effect is plausible enough, my theory is enough, I've got--Jerome Bruner voted for this. I should experiment.

I can't think of a single excellent publication that contains both thinking and numbers on that topic. Now let me give you the flip side of that. One way of framing that question is at what point do you think the program is
going to be a success and why? All right. And you then invest the money in the experiment.

Flip side of that is why did the program fail? All right. Do a lot of these trials or quasi-experiments, come up with estimates of effect that are close to zero or very small, and you then go off in a different direction. We in the social behavioral sciences, education sciences lack a kind of formal disciplined approach to doing postmortems, the thing went belly up. Right.

No significant effects or mixed effects or something. Understanding why that happened despite the fact that you had a lot of brain power going into this thing. The theory is good. You trust the people who are the service providers, the teachers, and so on and so forth. You trusted your graduate students who were already sworn to a vow of poverty, but are also exquisite at being conscientious record-keepers, testers and so on.

Why did the thing fail? We don't have tracks. We don't have books that sell at Brentano's or other places, that have titles similar to Why Bridges Fall Down, Why Buildings Collapse, Why the Cat Died.

This is another arena in which actually NSF could probably make a distinctive contribution. It's the flip side of your thing, understanding why it failed or why it could fail invites thinking about designing systems so that they have some redundancy.

What's a sufficient dose level? If we had postmortems on why the effect size was so small for just supplying the books, for example, that doesn't
seem actually to accord well with the way some people's theories would have it. So, there's lots of work to be done here. There's no easy answer certainly to your question, either your question, or the flip side of it, partly because nobody has done a lot of hard thinking and work on it.

MR. DELLA-PIANA: I have a question about fidelity and the worth of the experiments. In many cases, we know that good teachers would do fidelity plus. For example, in Kim's study, they would do leveled books, but they would make sure they had, as some of yours did, and were encouraged to have books that had a wide range of vocabulary, a greater range of discourse structure. Is this--how do you deal with this in your design?

Do you force people to do straight fidelity or do you make sure you're measuring, assessing the variation, and looking at those effects, or is it if you're finding that conceptually the program is one that a lot of experts or teachers feel is unethical because it's putting a lid on kids, do they reconceptualize the treatment? I guess I'm puzzled about that aspect of fidelity which might limit in what you do experimentally.

MR. BORUCH: That is a terribly important arena for both research and development. But I think it does vary depending on the character of the package and the intervention a lot so you have to sort of tailor the thinking to suit that.

Success for All and a couple of other packages are quite prescriptive. All right. In medical jargon, they call this manualization. You
manualize the hell out of everything, every step, every, you know, every time period, every--and this bores the socks off some people. They bridle.

The opposition’s claim is that the manualization suffocates creativity and so on and so forth. I've not seen Slavin's recent work on sort of measuring the extent to which that prescription, that manual is followed. The most current work on observing what's going on in that classroom, dimensionalizing it, understanding to what degree it's management, it what degree it's pedagogical, to what degree it is sort of dialectical dialogue in mutual education or learning.

Some of the most recent and best work is actually being supported by William T. Grant Foundation and, jeez, the guy who's doing it is actually at—I think at Virginia. I can't remember his name, whoops it's Paita. If you look at the annual report, just look in the Web site for William T. Grant Foundation, look at the grants awarded last year or in recent years for research on this topic, you can probably find him easily.

It's a good area to be in because, as I said, we don't have that sort of measurement technology, like IRTs for this context, and things like videotaping and understanding how, instead of videotaping, what are cheaper ways of doing it or how to sample, how to dimensionalize are all important, but sort of, as yet, under-researched.

MR. MARTINEZ: We are exactly on schedule right now so we're going to break for lunch. I want to thank our speaker, wonderful ideas and discussion.