MR. MARTINEZ: Let's reconvene for the last about 52 minutes. We have 52 minutes. The originally conceived purpose of this last hour is to have a discussion on the issues presented. So we would give the speakers time to question each other, for us, for our audience to question them, and I see Larry Suter is ready to go with the microphone. So that's what we want to do for this hour. That's the primary purpose.
There might be a second purpose also served in this hour which is to think about—well, I should back up and say this is the first of three workshops on experimentation and quasi-experimentation and attendant issues. We're planning to offer one in the Spring and the final one next fall.

We are open to topics, speakers, and we want people's ideas about optimal dates, too—second week in April or whatever works out for you. We'll do our best to maximize availability. I should say also that this is a fairly small—obviously. Strategically, we decided to start with MSP project staff.

I anticipate that as we look at the second and third workshops, we'll cast a wider net and issue invitations more broadly to other EHR programs, other EHR staff, and on we go from there. So, okay, and that takes us to your evaluations, the questions that I asked—future speakers, best available dates, future topics. Those are all questions that we want—we've placed on your evaluation forms, which before you leave today I hope you'll give to Janice Hansen who is going to raise her hand yet again.

In addition to that, if we have some time at the end of this hour, Jim Hamos suggested that we could spend some time actually discussing those questions about future topics and so on. So let's keep that open and maybe dedicate the last third or so of this hour to that.

Okay. Given that, the floor is pretty open right now, and I think Dr. Suter is ready to ask a question. Oh, one more thing— I'm sorry—for the sake of our transcriber, Victoria McLaughlin, there, we'll need to be sure to use the microphones. Okay.
MR. SUTER: Okay. Thank you, Mike. I was still thinking about that last session and I don't think Eamonn answered Boruch's question well enough. So I want to go back, I want to try one more time, and I'm going to put it in the context that we as program directors see a lot of proposals about education research, and I'm concerned about the quality of research, about the scientific merit, whatever that means, and if I listen to Eamonn, it means all this stuff, and if I listen to Bob, it means this box over here.

So I think we need some agreement here on what good science is in research and how we're going to get it and whether or not we can answer ultimate questions, what is good science?

MR. KELLY: Anything Bob publishes is good science. I mean I think it begs the question because one of the tensions of the National Science Foundation is that it's the National Science Foundation and it's not really an educational institution that does basic science, and that basic science should be free to roam wherever it wants to roam regardless of controversy.

When the findings from the National Science Foundation or science anywhere then become part of the public dialogue about what should be in schools and how to measure it, then you get into the controversies where you're moving from something which is primarily a hard science with pretty good models to understand it to hermeneutic field in which the issues have to be debated and negotiated as to what do you mean by the goals of science, how do you measure them, and how do you get there?
And I think that's a large, probably unresolvable issue because I think in one case, you're asking what is good scientific method assuming a declarative space of the hard sciences. What is a good scientific question in a clinical setting like schools or public health I think is a different, different set of assumptions, and the question is to what extent the models that pertain to the areas of the hard sciences port in some credible fashion?

And it may be the case that in some cases they do and some cases they don't. So the issue about what is scientific research, I think there's an awful lot to be said for work that brings new understanding. So we just had a meeting, a DRN; it was an EHR meeting, and the people in DUE talked about the concept inventory as a huge influence on undergraduate science education.

So there was a reconsideration of what does it mean for somebody to understand science and how do you measure that understanding of science? And that change in the construct of what it means to think like a scientist I think was and sometimes very scientific. It had a large impact without necessarily being related to effect sizes, and interesting, as you know from a review of the work done by Hestenes and others, it's very, very hard to change scores on the force concept inventory. It takes an awful lot of work.

Interesting how hard it is to have an effect. So there's a case in which we could do with some really good control trials to try to figure out how to drive up scores so that people leave Aristotle behind, which is where they live, and move into Newton, not to mention--

MR. BORUCH: Einstein.
MR. KELLY: Speaking of Einstein, do you have any comments to make?

MR. BORUCH: Oh, bless you, sir. Let me mention one. I think you brought up the matter of when do we know, when can you figure that an experiment might be justified or, you know, how big is the expectation of the size of the effect before you jump into this business?

Robert Yin just reminded me that there's another sort of answer implicit in other frameworks that says, at least in the medical sector, stage one, stage two, stage three. Stage one is sort of basically discovery. Stage two, safety, and stage three is the efficacy trial or the effectiveness trial.

Understanding how much time it takes at each stage sounds like a science policy variable. Understanding success rates at each stage for not only science stuff, but drugs or medicine, whatever, understanding what it costs to do each stage sounds like something a foundation like NSF could reasonably get at given its own access to its own sort of data archives on projects in the past.

With regard to the question, I think two people brought up the issue of experiments cost more. Well, certainly they take more in the sense of skilled human resources negotiation, a capacity, willingness to spend the time in developing the partnerships and so on.

Certainly that part takes more resources, and which could be expensive, but even in quasi-experiments or observational studies, somebody is out there trying to deploy the program, and so the randomization itself is, once
you figure out the design you want and so on, is relatively inexpensive part of the process.

Furthermore, when you look at the estimates of effects from the quasi-experimental designs as opposed to the randomized designs, typically after you've made, after you've done all this modeling, chewed up a bunch of degrees of freedom, it turns out that the variance of the estimates is bigger typically in the quasi-experiments.

So again I guess I think that this Foundation could and William T. Grant, they're actually in good positions to try to actually get good purchase on the idea of cost and whether the claims about large costs as opposed to small costs are justified.

MR. KELLY: And of course the costs are related to the cost to society of not doing the trial, getting a result and implementing it.

MR. BORUCH: Yeah. Incidentally, there's a lovely article by Fred Mosteller, Jimmy's mentor, one of Jimmy's mentors, and one of my dear departed colleagues, entitled something like "Cost Effectiveness of Randomized Trials in Medicine," written with an economist. It was an effort to try to understand can you justify this stuff in economic terms?

This is kind of shadowy arena, but the article is 25 years old, badly needs to be updated, probably ought to be updated somehow by folks tied in with the Foundation activity. And it would be fun to see that happen. Capitalizing on the economists in this arena is probably justified.
MR. KIM: I just have a brief comment about this question of what is good science? I think the answer to that is a little bit different in education, and many of you have probably read Donald Stokes' book Pasteur's Quadrant, but Stokes makes the distinction between research that just has basic knowledge as a goal and research that has use as a goal, use inspired research.

So you can think of basic research. The example he gives is Bohr's research on basic atomic structure and use-inspired research is Edison trying to develop the light bulb. Well, Pasteur is thinking about both, and he said, and one of the things that's very important about this book is that's basically what we're trying to do in education research.

I mean I think we're trying to advance both basic understanding of these human phenomenon, how do kids learn to read, how do they understand science, and we're trying to create research that is usable and useful to practitioners. So is it Larry--

MR. SUTER: Yes.

MR. KIM: So when Larry asked this question of what is good science, again, because I'm a very pragmatic thinker, I'll tell you the way I answer that when I think of undertaking studies. The way I answer that is I ask myself at the end of the day, after I finish a study, could I publish this in the Journal of Ed Psych or any other reputable peer-reviewed scientific journal that's very hard to publish in and could a superintendent, principal or teacher use this evidence?
If it only does the latter, all we're advancing is use-inspired knowledge along the lines of what Edison did that is really not linked to scientific theory; it doesn't advance basic understanding of human phenomenon. If all we do is more of the basic research, it's not as relevant.

So when I say that education research is a little different, I don't mean that causal inference is different or we don't use randomized field trial when questions of efficacy are asked. I just mean that we have a more difficult challenge, and so when you read a review of a proposal, I think that's an important question to ask, does it advance basic knowledge as well as use-inspired knowledge?

MR. HAMOS: I just want to sort of follow up and sort of get your thoughts on some of this sort of medical model and how it might apply to some of the work that I think is represented here.

So, in this audience, we have people that are practitioners, so people from the field that want to do this work and want to conduct good studies. You have many, many NSF program officers, but they were wondering what we funded. It was sort of Larry's question along the way.

We've never seen ourselves, as far as I know, as a portfolio of different approaches that will answer questions moving along the way to the big answer of what to do in STEM education or what are the things to do? Or we don't see multiple agencies even doing it that way.
So I'd like to get your thoughts on--and NSF also I think would never by itself because it's not our role to decide this is what everyone should be doing, and we never would--that we're really reflections of the field, and what the field thinks it should be doing, and then, you know, many, many people come in and out to be program officers and other things.

So what I'd like is sort of your thoughts on the medical model; should we create portfolios around days one, two, three, four? Be able to put different parts of our programs into those bins, help things move along the way? Is that a reasonable thing? Or should we create--is that just one construct and would you look at other constructs?

MR. BORUCH: NSF issued a report, maybe two years, I think. One of the principal authors was Barbara Schneider, and it was a review of the sort of joint ventures among the Department of Education and NSF in these IERI things. Okay.

I took a careful look through that, partly in the interest of discovering, well, are you folks doing any experiments, sponsoring any experiments at all? And it turns out you are, and also the view toward classify, stick into boxes stage one, two, three, or some other array, you know, six bullets, to try to get some sense of what was going on, given that it's curiosity-driven research from the field.

How does it--the first question is what's going on? Second question is can I categorize or can I sort of see the order in this chaos? All right. And frankly I think NSF shortchanges itself in actually failing to put things into
boxes in ways that the lay person, non-scientist, can appreciate. And also selling yourself—not that you're poor, but EHR is what--600 or 800 a year?

MR. SUTER: We're about 800 a year.

MR. BORUCH: Yeah, okay. But you could always use more.

[Laughter.]

MR. BORUCH: And understanding how to package the thing is good on sort of political grounds, but I think terribly important on intellectual grounds. I mean the thousand flowers blooming, the best flowers, of course, trying to figure out where the order is in this English garden as opposed to something trying to get the semblance of order sounds like a good thing to do, apart from the sort of real scientific benefit of understanding where the progress is being made, what the hit rate at each stage is, where the money is going for each stage.

If you look at, if you do a similar thing, to get back to the hard sciences stuff, look at the articles in Science magazine and try to classify them, addressing what kinds of questions, the vast majority are that first bullet, at least on my scale, what's the nature and character of the phenomenon, the rate change in this, the distance from that, the orders of magnitude for estimates—so a lot of scientific work is just like that, and the number of times you see formal tests of hypotheses in the Journal of Science is actually quite low.

All right. You do see them. You see confidence intervals, thank God, but it's that pig in Kevlar story.
MR. KELLY: Responding to the portfolio question, part of the issue is that there are so many variables that somebody from the field could look at. You know, you've got a developmental from baby, from cradle to gray, you have got all the STEM content, you've got all the context in which it could be done, you crossed that with technologies, you crossed that with a lot of different measurement devices, et cetera.

To allow that matrix to get peopled over time by accretion, you run into Cronbach's decade effect, you know, if the decade passes, the world has changed very radically and the interventions that worked a decade ago don't necessarily work now just because the passage of time.

You know, I mean judging, for example, the efficacy of, quote-unquote, "computers in education," when all you was a TRS-80 compared to Wikipedia and iPods and download, I mean the constructs change over decades, and it's very hard to do that.

One place where we're going to hopefully gain some data on that is that the new REESE has these targeted areas. You know the Foundation is doing a social experiment saying there are four or five areas that are really worth putting some money into, we think, based on emergent factors in other places.

And if that, if part of the portfolio can be put in, part of the money can be put in order to build the series of studies that would build upon one another. Otherwise, you're hoping for the Jim Kaput or the John Anderson who will do on their own phase one, two and three over 20 years. I mean the
Anderson stuff we saw is in the Clearinghouse, that's 20 years and about $20 million of funding including, not including money from ONR and maybe from the Department of Ed.

So what you've got is, if a researcher has it in mind that they want to do that, then you end up getting it, but if you trust the field all the time, part of the frustration with looking back at the time I was here in '97 to 2000, was that you get a genre of researchers that are addicted to novelty. You know they want to do the next new thing and the next new thing and they don't want to instantiate it and put it into curriculum and do a trial on it and then see how it diffuses.

You know there are some exciting new things coming down the pike so you get cycles of novelty that don't accrue and have some very good insights. I mean I believe the NSF—I mean, Mike, how much money has the NSF spent on curricula interventions over the last 25 years and how much of those are, you know, really good ideas and never got out into the field for whatever, I mean into schools, institutionalized?

MR. MARTINEZ: A response?

MR. KIM: I'll pass on this question.

MR. BORUCH: Actually, there's another—Eamonn's comment about sort of time is a variable in all this, it's one of those social science variables that we know we can measure more or less reliably, and cranking it into the conversations like this is terribly important.
Jimmy's response to Larry's question about what's good science included sort of this Pasteur thing and focusing on basic understandings, as well as the application side. Understanding what the shelf life of the knowledge is in this sector itself is one of those understudied and under thought about arenas.

We presume that the basic research is going to be durable. You know the speed of light is this thing, and we finally agreed on it, and it kicked all of the centers out.

On the other hand, on the social sector, understanding what that shelf life is for different kinds of interventions, for different kinds of outcome variables or different kinds of theories sounds like a good thing to look at. I've got no idea what, if I went back to the office and spent next week on it, you know, what would pop out thinking about this or reading about it or where I could find evidence on it.

But it does sound like if you're right, some of these, you know, ten years later, it's out of date--Cronbach's notorious pessimism about--or maybe not. Grouping, ability grouping versus not? Dropout versus not? There are an awful lot of things out that are fairly durable in their sort of persistence, if not for this country, then for other countries. Well--

MR. MARTINEZ: Other questions?

MR. KELLY: While you're waiting for an answer, you just reminded me, Mike's mentor, Bob Calfee, went down to Riverside and took everything that he knew about teaching reading and produced a very cheap
reading intervention, so cheap that you could do it on mimeograph machines, and actually ran them off on mimeograph machines, and got I think it was an IERI grant, to look at the impact of it, and the impact was dramatically positive.

He had taught reading for 30 years probably at Stanford, and then one day the principal came to him and says, well, we're stopping all that, and he said why? He said look at the data; the kids are reading. And she said, well, a decision was made by Sacramento; we're switching to a different reader, you know, different basal system so good-bye.

So there was a political decision which damaged the shelf life. I mean clearly there was something good going on there. He had learned something about how children read so the basic science may have been true, but it has to get enacted too, and sometimes that enactment is not a scientific variable.

MS. VANDERPUTTEN: This morning, there were two, well, we had three talks, but one talk, Jimmy, yours was very focused on a small intervention that could isolate to other interventions, and then Eamonn was talking very much about in order to understand the system, we have to have this full, huge contextual understanding.

So this workshop is part of the Math Science Partnership, which very largely was very large grants at trying to influence everything at once, and yet we also have some relatively small things within it. So I guess my question is what is rigor when you're trying to look at the system, when you're
trying to look at a targeted or a comprehensive math science partnership that is trying to bring about changes in many, many different ways, and you certainly can't randomize the whole thing? You might be able to randomize some parts of it. So what is rigor in large-scale implementation projects and how should we think about it?

MR. BORUCH: Systems level stuff.

MS. VANDERPUTTEN: Yeah.

MR. BORUCH: Good question.

MS. VANDERPUTTEN: Easy answer.

MR. BORUCH: Where is Robert Yin?

MS. VANDERPUTTEN: He's back there.

MR. BORUCH: Are people aware of the thing that you're trying to do with trying to marshal evidence and understand how it coheres or does not cohere in these large systems approaches to deploying intervention? Are you free to talk about that yet?

MR. YIN: Well, I'm free to talk about it. Like you said, the MSPs are large and obviously we've made multiple references to the award size. It does not necessarily mean they're singular, does not necessarily mean they are themselves systems-based. So the first copout for me is that in many large projects, in fact, the project that seems large may be a bundle of smaller projects.
Now, we're kind of moving back toward some of the methods and approaches that we've discussed, and within those smaller components, there is a legitimate question of, well, why isn't at least one of them a random trial of some sort? And are there many of them? And thus far the answer is there are not many of them.

Why there are not? I don't know. Because one could have taken some part or some aspect of your MSP and decided this part was worthy of investing in the design and ready, an intervention that was ready for some expected outcomes that could be tested through that design.

But if you go back further in NSF history, not too far back, there were systems projects, and those would be known, well, they were part of what was once an Office of Systemic Reform, and those would be ones known as State Systemic Initiatives or USPs or RSPs or the whole bunch of, maybe five or six different programs, all of which were trying to instigate systems change.

In those situations, I would have to start using some of Eamonn's slides and saying this is like was it the Irish or the English, you're not about to manipulate, or even Bob gave a technical constraint, which is, well, you've got one of these and one of those, even though you have many students in each school--this was his example of that you only randomly assigned two schools, so your systems research or systems interventions smack of the small "n" problem.

And you can't get around it even if you had lots of students because you have nesting and you all know those constraints. So there you'd have to
start using some other method. Now, I don't know if that's getting at where you wanted to go.

MR. KIM: I have two thoughts on this question of what does rigor look like in large-scale projects. I think on one level, this is just a statement that I would apply to a lot of concerns in education. You know our public school systems are governed by democratic institutions, school boards, and then they hire superintendents.

So to some extent those leaders of public schools should have some say in the kind of questions that are important for improving practice let's say in school districts. Well, when we go to school districts, when you actually talk to superintendents and principals and teachers, and you ask them what is an important question that needs to be answered to improve practice in reading or math or science, many of those questions happen to focus on the efficacy of particular programs.

So I think one criteria that I'd use in terms of what is rigor in public schooling and in terms of how we use evidence is that the decision-makers, the people who have the political power to stop effective reading programs, would kind of demand better evidence to enact programs before they're scaled up, that we would have--and I think that's when people make the analogy to medicine, that's usually what they're appealing to, that it's not based on ideology or politics, but there is some sort of empirical evidence that informs the decision.
The second thing that I would say about rigor is that I might have given you the misimpression that all I do is go around and randomly assign kids to different programs so I can do great experiments. I'm a historian by training. So this is very new to me, and one thing that I'd say about rigor is we're doing a study in southern California where it is a place-based randomized field trial where it's very complicated in that we're randomizing classrooms within schools to different experimental conditions.

Well, one obvious reason we're doing that is we want to know what's the effect of this cognitive training program for English teachers and their instruction and also student outcomes? Well, at the end of the study, we're going to make a comparison. It's going to be very complicated because we have multiple nesting issues and difficult instruments to work with, but I'm not just interested in does this program work?

I want to know what happened with instruction in those classrooms? So part of the rigor question is not just doing a horse race and figuring out which horse wins, but it's figuring out why. I mean what happens in those classrooms? What were some of the mechanisms? And that requires a lot of theory building and so forth.

So I think the danger of maybe emphasizing too much on randomized field trials is assuming that if we get evidence at the end of the day that shows that the intervention caused superior outcomes for students, that we're done, and I would say that we're not done; we have to figure out kind of the intervening reasons why that occurred.
And so in this particular study, we're really trying to get a lot of good implementation data. The one thing I'd say about getting the good implementation data is that there's a big tradeoff.

When you go in and start trying to get really good implementation data, where you go into the classroom and you observe the teachers and you survey them, there are potential big Hawthorne effects because now what you're doing is you're putting in the study all kinds of conditions that probably would not exist once you scaled this program up. So that's, again, that's a tradeoff.

But when we talk about rigor, let's be explicit about what those tradeoffs are, why you made this decision, and just point that, look, it's worth getting information on what's actually occurring in some of these classroom processes.

So I think we've got to think of rigor just beyond just a causal effect to what is actually happening that led to those outcomes.

MR. MARTINEZ: I want to follow up on that question, too. I guess I wonder if there might be a couple of other ways to think about rigor. One is through a framework, a validity framework. So we've talked about, for example, construct validity. We have focused largely on inferences about causality and moreover there's a conceptual area of external validity and statistical validity.

So maybe on the conceptual level, a validity framework can help us to think about rigor and to sort of get a handle on level of validity according
to established typologies, and on the other side, there might be technical aspects of rigor that we've sort of touched on today, but really haven't explored very deeply.

Now I'm not an expert in this area, but I would say things like multiple pretest, kind of baseline pretest, matching or stratification, latent ability scores using multiple measures, quantifying implementation. That is dosage so that we can really dosage to outcomes.

Dr. Boruch mentioned intent to treat. Thinking about confounding variables most likely to influence outcomes and measuring of those, going after those directly. So I wonder if part of the answer might be technical design elements of the study that we really haven't had time to delve in today, but which we could go into more deeply in future meetings.

MR. KELLY: It would appear that the question of rigor applies regardless of the particular methodology one uses, you know, there's a lot of debate going on in anthropology about what makes a good ethnography and how far do you push the assumptions? So it would appear that whatever method somebody was using, was videotaping or surveys or randomized trials or whatever, they should be doing it in accord with the current best negotiated standards for quality within the area that they're deriving the methods from because--and I think the NRC book is right on this point, that part of doing good science is being able to rule out, you know, alternative explanations and to have some sense of--it depends on the question, of course, and on the claim, but the questions of prediction and applicability of theories and so forth.
I mean there's a number of different criteria that you could come up with for each one of the methods, and it seems that if somebody is making a claim that they are using a method that gathers the evidence that allows them to make that claim, based on that structured abstract, you know that fulfills other criteria by Toulmin, you know, that it's not subject to rebuttal, that you admit your qualifiers and so forth, so that no one study is necessarily more rigorous than another, and no one study is going to be a final point. It's going to be part of an ongoing renegotiation, hopefully across many groups.

I mean you're talking about the implementation question is a very important question, and not just the implementation, but the understanding by the teachers of what it is that they think that they're teaching and the understanding by the students of what it is that they think they're getting.

You know one of the people we fund here is Joanne Lobato who is making a career out of what she calls actor-oriented transfer models. In the transfer of learning framework, going back to Thorndyke, you've got two tasks that are similar to the researcher, and you train on one, and then the person is supposed to transfer that learning on to the second task. And there's hundreds and hundreds of studies to show that that happens rarely.

So you've got this odd phenomenon where you can't show people are learning, but you know they're learning all the time; right? But how do you catch it in the laboratory? And as I said, there was a special issue in the Journal of Learning Sciences on that.
Her take on it was to think about it from the learner's point of view. So she would take the text and say to the learner, what do you see in this text? And it may not be at all what the researcher had seen in the text. Okay. What do you see in the second text? Again, nothing that the researcher had thought was important, but the student is seeing a similarity.

And that is how she explains a lot of transfer that is going on. In the former method, that would be discarded as negative transfer or something kind of bad or a failure, to use Bob's earlier comment. But from her point of view, a student coming to know is a novice, and how they look at the world and build it up is something that the now expert had to do developmentally also.

So because an expert can see two things in these tests, and the novice doesn't see it, doesn't necessarily mean it's a failure; right? It may be a developmental thing, and this is kind of the insight from Hestines and others, which of course goes back to the idea of evidence and validity regardless of which research method you use; right.

MR. BORUCH: You're bringing up ethnographers, that Shavelson and Towne Academy committee invited a fair number of different kinds of people, different characters, to sort of provide testimony, including some anthropologists who basically declared that, look, if somebody can replicate my results, it means I have not exercised enough creativity, originality and the like in producing these results. If I'm replicable, what good am I? What's my value added; right?
Now, that's a perspective. It is a reasonable perspective from a variety of points of view, but understanding where it fits if at all into a scientific perspective, and replicability is part and parcel of what we think about, led to our kind of ignoring that particular branch of anthropology, ethnography in the volume.

The second piece to this has to do with the extent to which there are developing standards for more process-oriented activity, monitoring, looking at--this is the Yogi Berra, if you want to find out, if you want to see what's going on, go around and look and see what's going on--Yogi Berra approach to science.

Among others, the UK Cabinet Office, which had a Department of Research and Evaluation until recently--it's moved over to Treasury, I think--let contracts to develop standards for the qualitated area because there was sufficient interest in that country and sufficient expertise, sufficient skiddishness and worrying about, well, how do we know we are doing good work that may produce some interesting tracks on judging the defensibility of the evidence, the warrant for the claims, enunciated by the anthropologists.

Historically, Ozzie Werner--do you remember that name from ancient anthropology--this is Northwestern 25 years ago--and others produced anthropology text in which in the appendices they provided information about how to judge the defensibility of the claims made by the anthropologists, the nature of corroboration of evidence in the context of anthropology.
It would be nice to see something that's similar for things like what comes about as a consequence of videotaping or observing, surveilling stuff in the classrooms. We don't seem to have it yet or at least I haven't seen it. Maybe you guys are inventing it as you go.

Oh, one minor note on the surveillance, I see that surveillance, that is the extent to which people change their behavior because you're watching them so closely, for years, criminologists knew that, both criminology research and social welfare research, the experiments themselves and the measurement processes that they involved did induce higher rates of crime, detecting crime; therefore, it made crime rates look like they were going up in some of these experiments rather than down, as they were supposed to be.

Child abuse and neglect went up because you were surveilling people a little bit more closely. You could see more of it. So it's nontrivial. It gets a little bit more, even more nuanced than the Hawthorne stuff, but it's, at least in those arenas, it's, I think it's at least as tough as it is in education.

MR. KELLY: I must say I've always been annoyed by that Yogi Berra comment in the book. I'll see your Yogi Berra and raise you a Buckminster Fuller, who once said it's very hard to teach logic in a world where people say the sun is rising when actually the horizon is falling, if you just look.

[Laughter.]
MR. KELLY: As for the replicability, the replicability is an odd thing because it goes back again to values; right. I mean the phenomenon should be replicable; in some sense, maybe it shouldn't. There's a field called visual analytics in homeland security where they are trying to find, they're trying to discover that which by design will be nonreplicable, which is sniffing out terrorist activity.

You know the chances that you're going to get a repeat, a replication of a particular terrorist act gets smaller after it's been done. So visual analytics is the combination of visualization software together with data streams on cell phones and Internet traffic, what's coming in at the ports, in order to create a science of suspicion, right, to anticipate something which will probably be a once off thing and stop it, which is really a hard task.

You know they talk about people joining the dots, the number of lines it takes to join six dots is actually quite high. If you just had six dots, and you're talking about dots you don't even know exists. So I mean it would be interesting to see how something which is--I hate to think that that's ruled out of science, this kind of activity.

MR. BORUCH: On the terrorism stuff, those guys, everybody has got problems in these various domains. Campbell Colloboration's Crime and Justice Group had a couple of people interested in generating a systematic review of evidence on the effectiveness of antiterrorism programs. Two of them were actually at George Mason University--Cynthia Lum, and I think--do you have a criminology department or she may be in sociology.
And the exercise, the business of searching through both published literature and as much gray, unclassified gray literature as one could get a-hold of, suggested strongly that the evidence being accumulated in this arena is actually pretty sparse, pretty weak.

With respect to sort of replicability, one of the main outcome variables here is repeat events. That is as soon as you initiate, let's say you do a retaliatory raid, it's supposed to be in some sense counterterrorist, right, supposed to reduce the likelihood of a subsequent event, if instead it enhances the likelihood that another event occurs close by, close proximity in time, you've got a negative effect as opposed to a positive effect.

So this pattern of recognition stuff, how many lines you need for the six dot problem is different, no less interesting, but the fact that even in that sector, which is awash with money, driven by fears that are sometimes just outlandish, those guys have won in the sense of our redirecting so much cash into the counterterrorism bucket, it's not even funny, but--

MR. SUTER: But are they doing good science? You say they're not?

MR. BORUCH: At this stage, the best of the--well, it's hard to do control trials in that arena, too, you know.

[Laughter.]

MR. BORUCH: But when you take a look at the best, the best available data is typically a time series on body counts, episodes, events,
volatile, you know, engagements, all the rest of that stuff, and you can do as
much in the way of drumming up interesting comparisons regionally or cross
regionally, but it's still the case that an awful lot of these things, there's just
not very much evidence on.

Including, for example, estimating how many people, well, how
many terrorism activities have you actually forestalled or prevented as a
consequence of all that waiting in line at the airport for stuff and so on and so
forth? Could you accomplish the same thing with a couple of dogs walking
around as you could with that armamentarium that they afflict us with at every
airport?

MR. KIM: So I'm not going to talk about the replication point as
much as how about in terms of good science correct interpretation of results?
So one of the basic lessons that we all learned in Statistics 101 is that
interaction effects render main effects irrelevant.

In other words, if you find that a treatment has a larger effect on,
say, poor kids, you're not going to then infer that this intervention has
positive effects on all kids. So this is actually a really important point I think
for scientists to understand because sometimes we think we do the great trial
with these great results, and then it will kind of take care of itself, then we'll
have good policy.

But, in fact, that doesn't happen often, and I think the best
example of this, of the contingencies under which it does or does not happen, is
what happened after the Tennessee class size experiment that Bob and I and
others, you know, we talk about this being the gold standard in education. Fred Mosteller said this is the kind of study that we need. And, of course, California takes the results and it reduces class sizes in all of its schools K through 3.

So everyone gets class size reduction, but what did that study actually find? It found an interaction. In fact, it found that the largest effects were for inner-city minority school children. Well, what did Tennessee do when it found those results? Ned McWherter, the governor of Tennessee in 1989—he's a Democrat—this study was supported through a Republican administration, Governor Alexander, and McWherter basically looks at the results, and he says, look, the impacts are largest for minority inner-city school children; we need to target this intervention.

So he said to the 139 superintendents in the state of Tennessee, I challenge you to cobble together your Federal Title I dollars with your local tax dollars to reduce class sizes, and I will find state discretionary funds to match those dollars.

As a result, the state reduced class sizes in the 16 highest poverty districts. And what it found over the next four years is these 16 districts relative to the 139 were at the bottom quartile in achievement, and five years later, they had moved up to about the median. Can't draw a causal inference, but it's certainly correlational.

So what's my point in saying this? It goes back again to the danger of focusing so much on randomized field trials as the gold standard is that we forget it's the gold standard because it's the means to this end. It's the means
to clearly communicating results that can really inform policy and practice, and in our multi-racial democracy address these kind of pressing social concerns.

But for that leap to happen, from the end of the randomized field trial to the use of that evidence, I think scientists, folks like us, actually have to be much more in the policy arena or else we get I think misapplications of policy as we did in California, rather than something like Tennessee.

MR. KELLY: I believe it's the case that in the California system, it's not like there were thousands of good teachers just waiting to get hired.

MR. KIM: Right.

MR. KELLY: They hired people kind of off the street, and a lot of those kids, you got reduction in school size, but they didn't get much instruction because the people weren't certified to teach in the first place which was another sad outcome of it.

Actually I'd be interested to hear from both of you in terms of your response to Cronbach's contention that the number of interactions grows to the point where you can't communicate anything. I mean once you get interaction of three terms or more, so you say to the superintendent, if they're poor and they're this and they're that, and it's four o'clock in the afternoon, then try this.

But if they're--I mean you get to a point as he said with these hall of mirrors example, right, where you end up so overqualifying your statements, that you don't have main effects, you don't actually have second order effects,
you may not even have third order effects, and as you increase the number of variables, given the complexity of the phenomena you're trying to look at, you're probably going to find more and more interaction.

So you come to the point where you have to caveat it to the point where you may not be able to say something unless you do an absolutely against a universal experiment and do Laplace's "demon" to tell you what to do next.

I mean how do you guys respond to Cronbach's challenge? When was that, Mike? That was--

MR. MARTINEZ: That was 1975.

MR. KELLY: 1975.

MR. MARTINEZ: And if you want to answer that question, you have 30 seconds to do so.

[Laughter.]

MR. KIM: Can I try to do that in 30 seconds? Here's my 30-second response. One is everything in education is ready-made for meta-analysis. Look, the first thing we learn as social scientists is you never base a policy on one experiment. It's not Tennessee. You got to think about Connecticut where they found no effects, and look at the distribution to see what all of these studies are saying cumulatively.

The second response I would have is you've got to, your interactions have to flow from some theory and your explanation of why class
size effects are largest with little kids has to flow from a theory, and Ed Lazear's work—he's an economist at Stanford who's developed some of these theories about how younger children misbehave more and so there are more disruptions, whereas with older children, let's say in a graduate school lecture of 100 kids, there's no misbehavior and therefore there is more efficiency in learning.

So you got to think those two things I think are pretty critical. One is the whole meta-analysis issue shedding light on the effects that you're getting to kind of put it in context; and two, your interactions or main effects have to be explained and understood in terms of some theory, which I believe Eamonn is pretty skeptical about, from what I could tell. That's my response.

MR. MARTINEZ: Good. He can't answer.

[Laughter.]

MR. KELLY: I'm not saying anything.

MR. MARTINEZ: Last comment. It's yours.

MR. BORUCH: I agree with the preceding comments. The only thing I would add on top of the fact that those three and four-way interactions are actually (a) typically impossible to estimate given the resources you got; and given the resources you got, if you try to estimate them, the variances are so high that nothing turns out to be, quote, "significant."
Looking, at least, for the two ways that are important, on policy grounds and on theoretical grounds, the two may coincide halfway at times, sounds like the most sort of practical approach to getting a handle on this, but make sure that more than one person does it so you get the replications and the meta-analysis.

MR. KELLY: Hopefully, in terms of replications, we live in a two-factor interacting world.

[Laughter.]

MR. BORUCH: Ceiling on number of interactions allowed in this room.

MR. KELLY: Or else we're being fooled.

MR. MARTINEZ: Well, I think I let Jim Hamos down because he really wanted to talk about next steps and so on, but there are a couple of ways to do that. You have your evaluation forms and we really do want your thoughts. It's not about--well, as you know and as we've said, we're planning two additional workshops, so timing, speakers, topics, very, very important.

And beyond, if you want to discuss, stay around and discuss, besides filling out the evaluation form, Janice, I, perhaps Elizabeth, I don't know, others, will be around for a little while, can stay for a bit.

I really want to thank our speakers for being very, very provocative today, making us think, we have some answers, maybe a lot more
questions, but in my view, they set the perfect tone for what we wanted to do--what I had hoped to accomplish today, and we'll look forward to rolling this out in the future. So thank you very much, Dr. Kim, Dr. Kelly, Dr. Boruch, fantastic.

And in addition, I want to thank the National Science Foundation for supporting this project, and we're working on posting materials from today and transcripts on MSPNet. Not sure when they're going to be up, but we're working on that.

Two other people--

MS. VANDERPUTTEN: If any of you are not on MSPNet and want to be on it, let me know.

MR. MARTINEZ: Very good. Two other people I want to thank, my graduate student Janice Hansen over there, who is phenomenal, and she's in our first crop of Ph.D. students at UC Irvine Department of Education--she's on the dream team, she reminds me; she's absolutely right--and my program officer, Elizabeth VanderPutten, thank you very much.

It's been fantastic, and I should mention one other person, Eamonn Kelly, who is actually my advisory board, sort of one-person advisory board. So thanks for being here, and have a wonderful afternoon. See you in the Spring.

[Applause.]

[Whereupon, at 4:00 p.m., the workshop was adjourned.]