INSTITUTE OF EDUCATION SCIENCES

FIFTH ANNUAL IES RESEARCH CONFERENCE

CONNECTING RESEARCH, POLICY AND PRACTICE

CONCURRENT PANEL SESSION I

"Beyond p-Values: Characterizing Education Intervention Effects in Meaningful Ways"

Tuesday, June 29, 2010

10:20 a.m.

The Gaylord at National Harbor
National Harbor 12-13
201 Waterfront Street
National Harbor, Maryland 20745

C O N T E N T S

- - -

P R O C E E D I N G S

DR. OKAGAKI: Good morning. There are seats up in the front rows for all of you who are walking in right now, and if you can hurry up and find a seat, we're going to get started.

While the last people are getting settled, I want to remind you that this session is being recorded, so would you please silence all of your electronic devices.

My name is Lynn Okagaki. I am the Commissioner for Education Research at IES, and this morning, it is my great pleasure to introduce to you Mark Lipsey. Mark is the Director of the Peabody Research Institute and Research Professor at Peabody College of Education and Human Development at Vanderbilt University.

Since IES began, which is when I started, I came to Washington, D.C., I've come to know and really appreciate Mark Lipsey. His CV is incredibly lengthy, and I'm not going to tell you everything about him, but I think that you learn something about people from the titles of the papers that

they've written.

So, for example, in 1997, Mark published an article entitled "What Can You Build with Thousands of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation."

In 2003, perhaps my favorite title: "Those Confounded Moderators in Meta-Analysis: Good, Bad, Ugly."

And then what is perhaps most relevant for today's talk, he did a paper in 2005 called "The Challenges of Interpreting Research for Use by Practitioners."

Today's session is about how do we communicate our statistical findings in ways that will make sense to practitioners, you know, p-values, effect sizes, those things that help us get articles published in good journals? They mean very little to the practice community. About a year ago, Mark and a group of his graduate students started working on this project, and we're looking forward to what they've learned about trying to communicate research findings to practitioners.

Please join me in welcoming Mark Lipsey.

[Applause.]

DR. LIPSEY: Thank you, Lynn.

A little bit of a preface here. First, this presentation is highlights from a paper that IES has commissioned that's in draft form, and should move forward through the review process and be generally available at some point, and has much more information in it, of course, than I can present here.

Secondly, let me acknowledge my collaborators, all of whom are listed here, a mix of graduate students from our elite IES-funded training program and some of our fine staff at the Peabody Research Institute, some of whom are probably in the room if I can actually see who is in the room given these lights.

Also, note at the bottom, I mention Howard Bloom, Carolyn Hill, and Alison Black, who collaborated with me on an earlier project about practical significance of effect sizes, some of which, some of that material is adapted also for

these purposes.

So, let's start out at the very beginning here. We're talking about intervention research, and the intervention research paradigm here is a familiar one to all of you. We're going to compare a treatment sample with a control sample, configured in some way—we tend to favor random assignment for good reasons—on some educational outcome measure as a way of estimating the effect of that intervention on that outcome measure.

What we get out of that research paradigm in terms of our sort of native empirical statistical findings are the means for that outcome measure on the treatment group, the control group, the difference between the means, and a p-value—that precious p-value that's the first thing we all look at for the statistical significance of the difference between those means. And that's the basic analytic finding.

Now, my point here, and sort of the point of the presentation, is really communication, presentation, representation of these things. These

native statistical findings do represent the effect of an education outcome, but they provide very little insight into what that effect actually means.

Practitioners, policymakers, and, frankly, most researchers don't really have a good framework or a good understanding of interpreting some means and p-values and telling you what that actually means in any practical terms or any intuitive terms.

A simple example. Let's say we've got a vocabulary-building program. We've got fifth graders in a treatment group and a control group. Our outcome is the CAT5, California Achievement Test, Edition 5, Reading Achievement Test. Our findings are that the mean score for the treatment group is 718 on that outcome measure. The mean score for the control group is 703. The difference, as you can see, is 15 points, and, by gosh, it is statistically significant. P is less than .05. Okay. That's what we find from the research study.

Note my little note there in yellow,

incidentally, because there's lots of places where I'm going to be tempted to digress, and this is one of them, p-values, just to remind you, are often touted as if they're indicators of the magnitude of an intervention effect. They are, of course, no such thing. They are heavily driven by standard errors and sample size. You can have exactly the same intervention effect that's nonsignificant, significant, and highly significant, and very, very, very significant, as we sometimes like to say in our articles, and it's all the same effect size. It's really a function of the power and the sample size.

We don't want to get into marginally significant, very significant, and highly significant, as if that really tells us something about the magnitude of the effect.

So here's the question for you: 15-point difference on the CAT5—is that a big effect or is that a trivial one? The students read a whole lot better than they did before or just a little bit better? If they were poor readers before, is that

enough of a change to make them proficient readers now? If they were behind their peers, have they now caught up?

Some of you who may be intimately familiar with the CAT5, what the Reading Achievement Scale measures, how it's scored and normed, might have answers to these questions simply from the statistical information we get out of our study. But I think most of us are really pretty clueless as to how to answer these questions—for a 15-point difference out of a—on a 700-point scale in relatively arbitrary scale units, for what that actually means.

So, that's the problem here. And what I want to do is just sort of overview two approaches to representing or presenting our intervention effects in ways that make it easier for us to understand them and certainly for practitioners and policymakers and others to at least have some comprehension beyond our means and p-values as to what it is we found.

The first thing I want to review is some

approaches simply to represent these findings in different forms descriptively. So, basically, to translate from the native statistical form into some other language or some other picture or some other representation that is intuitively more appealing and more easier to understand.

And there are, in fact, a number of quite easy ways to do this that we could do routinely that would be helpful, I think, in communicating and actually appreciating the results of our intervention.

The other part of what I want to talk about is a little bit of a reprise on something I've talked about here at the IES conference before, which is assessing the practical significance of these findings. So you've got descriptions-more intuitive descriptions, less intuitive descriptions but then you've got practical significance, assessing the practical magnitude of that.

I'm going to argue that to get to practical significance, we need something more than

just good descriptions of the effect sizes. We need some externally derived standard or criteria for assessing what is practically significant in the context, and I'll show you some examples of ways we might think about doing that more routinely in our research.

To start with part one, "Useful Descriptive Representations of Intervention Effects," we have to acknowledge at the very beginning that there are some cases where representation in terms of the original metric is really relatively satisfactory because some of these metrics in education are inherently meaningful.

So, if we're looking at outcomes that have to do with proportion of days a student was absent or the proportion who graduated, number of suspensions or expulsions, proportion of assignments completed-things on those order, particular events that are familiar and recognizable in educational context, and we're proportioning them or counting them that doesn't

present any particular problem.

The problems, of course, come in with things like achievement scales and attitude scales and other indexes that are basically in arbitrary units that don't have any clear meaning or any familiarity.

A couple of sort of sidebar notes here while we're talking about the original metric before we move on to some translations.

We need to think about covariate adjusted means for that original metric for much of what we want to do. Whatever the research design, there's always the possibility of some baseline differences that are going to unfold forward to produce differences at the end. Even in a random assignment study, there are often some baseline differences, and, of course, in nonrandomized studies, the problem may be worse. Small sample randomization can produce very large baseline differences.

Those are not part of the treatment effect. If there's a difference at the beginning that carries forward to the end, that's not part of

the intervention effect, obviously.

We also, of course, recognize problems of attrition. We may start out with very equivalent results, very equivalent groups, but if there's attrition, then what we end up with at the posttest may not be so equivalent. The standard approach is to covariate-adjust for those baseline differences as much as possible.

What we're looking for here when we're thinking about our original metric, of course, is our best estimate of the treatment effect, and that's usually going to be some adjusted or covariate adjusted values, not the original values. We want to keep that in mind.

Also, while we're talking about the original metric, I want to make a pitch for paying attention to pretests in a different way than we typically do. They're often used as covariates. We frequently compare pretest values for the two groups to see if we've got good equivalence between the—good initial equivalence between the groups.

But we can also look at pretest/posttest

change for each of the arms of our study: the treatment group and the control group. And that provides an interesting context just around our original metric for interpreting the nature of our effects, and let me show you an example here.

Here's some middle school students in a conflict resolution program, and we have some composite measure of interpersonal aggression. We've done surveys at the beginning of the school year and the end of the school year. Beginning of the school year is our baseline. Program is implemented in some of the schools, not in others. We do the surveys at the end, and here are the results we get then in three different scenarios.

Arbitrary scale here for amount of interpersonal regression—aggression—they may be regressing, too—but interpersonal aggression that the students report in these surveys. Notice in all three of my scenarios that the posttest results are exactly the same. Okay. The effects as measured at posttest are exactly the same.

Notice also that since I manufactured this

data, it's beautifully equivalent on the pretest, so we start out with equivalent groups in all cases, but the baseline values are at quite different levels, as you can see.

Now, if we look at pre/post change, each of these scenarios is giving us a different picture of the nature of this intervention effect. In Scenario One, we've got a slight decrease in aggression in the treatment group and a slight increase in aggression in the control group, and that produces the difference we see at posttest.

Scenario two, on the other hand, everybody gets worse by the end of the year. They just get less worse if they had the intervention. All right? Do you see that?

Scenario three, on the other hand, the control group gets a lot worse, a lot more aggressive. The treatment group does not change that much. What the treatment did in this case was basically stabilize the baseline value, didn't get much worse, didn't get much better, wasn't actually that much change in the intervention group, but by

comparison with the control group, which was getting a whole lot worse, you got a better outcome.

Do you see how this simple picture gives us a different image of what this actual treatment effect is all about simply by seeing where we started and what the intervention is actually doing?

Another implication of this, if you want to press this a little further, is that we can look at this posttest difference in relationship to the amount of change between pretest and posttest that we see, for example, in the control group. So, in Scenario B, for instance, where everybody is getting worse by quite a bit, the effect of the intervention to ameliorate or mitigate that bad trend is fairly modest by comparison to how much change there is from beginning to end.

Whereas, in Scenario A, for instance, where there's not a whole lot of change in the control group, it gets a little bit worse, and the treatment group gets a little bit better, but the

difference at the end compared to the little bit worse that the control got is really quite large.

You see what I'm doing here? Relative to the change, relative to what would have happened otherwise, what does that treatment effect look like? Did it change a whole lot—the pre/post difference? Change it a little? And in which direction?

All I've done here is to take my original metric, pay closer attention to the pretest, the change from baseline to the end, and use that as part of the context for interpreting what that posttest treatment effect looks like.

Let's move on to the granddaddy here of trying to represent treatment effects, and that's your effect size. I say here that that's typically the standardized mean difference just as a reminder that there's more than one effect size out there. There are lots of different effect-size statistics.

When we say "effect size," we're usually thinking about the standardized mean difference, which is shown here, which is the ratio of the

difference between the mean for the treatment group and the mean for the control group divided by what's presumptively the common standard deviation in a homogeneity of variance situation.

For those of you who are looking carefully in order to kind of make this picture clear, there are about three standard deviations between the mean of the treatment group and the mean of the control group. You should live so long as to have a treatment effect like that.

[Laughter.]

DR. LIPSEY: If I made it closer to what we actually find in education, you'd barely notice that these two distributions were different, and it would kind of mess up the picture.

Effect sizes. We're moving rapidly in most fields to a convention of reporting effect sizes along with our p-values and means and so on, and I think that's basically a good thing, particularly if we put confidence intervals around them.

But it's not much of a solution to the problem I'm dealing with here of making

intervention effects sort of understandable at a more intuitive level. These effect sizes—as I know well, as someone who does a lot of meta-analysis—are very useful for comparing effects across studies and across outcome measures because of the standardization there.

And they're somewhat meaningful to researchers. If you work with a lot of effect sizes for a particular kind of outcome in a particular intervention area, you eventually get some sense of sort of what's a big one and what's a small one and what they might mean, but this is not obvious. You go out to your typical teacher or policymaker and say, "I've got an effect size of .25 standard deviations on an achievement measure," and see how quickly they grasp exactly what it is, what the importance of that finding is.

So, it's not very intuitive. It provides relatively little insight into the actual nature and magnitude of the effect, especially for nonresearchers who aren't familiar with it.

It needs some kind of a reference to be

meaningful, and so they're often reported in relationship to Cohen's guidelines for what's a small effect size, a medium effect size, and a large effect size, and I suspect everybody in this room has encountered these in one form or another.

If there's anything I want you to take away from this talk is that that's a really bad idea. Okay? I want you to put your hand over your heart and swear that you will never mention Cohen's small, medium, and large guidelines in the context of educational research ever again, and I'll show you why a little bit later.

But the gist of it is that these are very broad summaries of what's believed to be out there empirically with all kinds of different measures in all kinds of different interventions and all kinds of different disciplines, and its applicability to any given measure in any given intervention study is highly questionable, and I'll show you that they're particularly bad for our most common educational outcomes later on.

Just a few quirks about the effect size

while we're talking about it, and then we will move on. Covariate adjusted means. Again, when you're computing these things, in your numerator, you want your best estimate of the treatment effect, and covariate means are relatively easy to get for most of our statistical analyses that have covariates in it.

The dummy code—the regression coefficient on the dummy coded treatment—provides that or, in an analysis, a covariance format. Most statistical programs will give you the covariate adjusted means.

You don't want to mess with the denominator, however. Okay? You don't adjust the variance or the standard deviation in the denominator of that. Something that's worth highlighting here while we're talking about effect sizes, the concept of the effect size—the essential character of the effect size—is that it's trying to standardize the difference between two groups.

Standardization means it needs to be done the same way in a consistent fashion every time,

and it's standardizing against the individual variation presumed to be a sample estimate of the individual variation presumed to be in the population.

I find that this standardization, that the lack of awareness about this standardization function confuses a lot of people who get into, say, multilevel models and are wondering what the effect size is if you've got a multilevel model. Well, it's the same thing it always is. It's the difference between the covariate adjusted means divided by the unadjusted variation, and you're done. There's nothing particularly fancy about that except how you get the variance components out of the model if you don't want to just calculate them without the model.

There is a complication in education with effect sizes which has to do with the variance on which you're going to standardize, and this is another area where we have some confusion. All the more reason why effect sizes are not particularly communicative unless you know a whole lot about

them.

But we have variance components in the world of education. You're all familiar with nesting of students within classrooms and classrooms within schools, and why do we care about that? It's, well, because students aren't randomly assigned to states, districts, classrooms, schools, and so on and so forth. They tend to be more homogeneous within and vary between, and that means you've got this complicated variance structure.

So, if we think about the total variance, national variance on achievement measures, say, for students, that could be decomposed into the variance component that represented difference between states, difference between districts, difference between schools, difference between classrooms, and finally students within classrooms, within schools, within districts, within states.

Well, depending on what kind of a sample you have, when you create an effect size, you've got different variance components in there. If you're working with NAEP data, and you have a

national sample, and you produce an effect size, that's an effect size that is standardized on what I'm calling Sigma squared total here.

If you've got data from a single school or a single classroom, you've got one of these other variance components, and you have eliminated some. Now if these variance components were all of trivial size, this wouldn't much matter. You'd get pretty much the effect size, but we know, and one of the reasons we have to be careful about multilevel modeling, is that these are often not of trivial size.

For example, we know that the variance between schools, if we think of the ICCs—the intercluster correlation coefficients—it's about 20 percent of the total variance. You take the square root of that in these effect-size calculations, and you change the effect size a good bit. So, our standardization of effect sizes in education is pretty imperfect because we're not paying attention to whether or not different studies are actually standardizing on the same set of variance

components, which further complicates the picture here.

Another complication that I just want to highlight as we're going here is that you occasionally run into people who are standardizing on a different variance and getting a different kind of effect size.

Now, there is nothing inherently wrong with that, but it's not always recognized that it's different. So, for instance, suppose I aggregate my achievement data up to the school level, and I have a mean for each school, and I've got some treatment schools and some control schools, and I take that aggregate data at the school level, and I compute an effect size.

Now, I've got the difference between the means for the treatment schools and the control schools, and that's pretty much what I would get no matter how I calculated that, but the mean I've aggregated up to the school level now, that's the variance component in here that's called "schools." And it's about one-fifth as large as the total or

something on that order.

If I standardize on that as a denominator, I'm going to have an effect size, but it's an effect size standardized in a different way than the conventional level. It's going to be much bigger.

I've got a small collection of papers that amuse me and distress me in some ways, in which that effect size has been produced clearly. No comment is made about the fact that it's not the usual effect size, and then what's often even more interesting is that it's compared with Cohen's standards, and this effect size is 1.25, and anybody can see that that's absolutely huge because Cohen says anything over .80 is really a large effect size.

Well, no, actually, Cohen's standards apply to variation between individuals. In our case it is variation between subjects, not variation between schools. Are you following me?

This effect-size business is a little more complicated than it looks like on the surface. The

standardization is important. It allows us to compare across outcome variables, across studies. To be of any use, it has to be standardized in the same way. It often isn't standardized in the same way, and even figuring out how it should be standardized can challenge a researcher and, above all, when we're done, it's not going to communicate very well unless you're very much into effect sizes for that particular construct what the meaning of the intervention effect is.

I mentioned, another of my sidebars, I already mentioned the multilevel analysis results. One of the most common questions I get as somebody who works with effect sizes is, well, how do I get an effect size out of HLM? So I just want to mention again, the same way you get an effect size everywhere else—difference between the covariate adjusted means divided by the unadjusted variation at the individual subject level.

So, now let's move on to some things that we might do that would make our intervention effects a little more understandable to us as well

as to whatever consumers we have, and one of the approaches that is easiest and that I like best is to translate our findings of mean differences into proportions—proportion of the treatment group that's in some category versus proportion of the control group.

I'm not talking about doing the analysis that way. We don't want to do simple but crude dichotomizations on continuous data because that has all kinds of adverse effects on our statistical analysis. I'm just talking about translating after it's done.

Here are our two distributions. This is just my effect-size picture again. We can pick a score anywhere along the continuum that is interesting or meaningful to us, and we can compute quite easily from our empirical data or, if we're so inclined, with normal distributions, turn everything into z-scores and go to a normal table and pull out the areas under the curve.

We can compute the proportion of the treatment group and the control group above and

beyond that threshold value. So, now we've translated our difference in means into a proportion. This proportion of the kids were successful by this standard, and the intervention pushed how many more proportionately of the students above that threshold?

And there are a number of sort of default options out there in the literature for picking a threshold. I actually think in most cases we can do better than that, but one of the most well known is what Cohen called by the memorable name "U3." There's a U1, a U2, and a U4 that nobody is interested in, kind of like the guy who invented one-up, two-up, three-up, four-up, five-up, six-up, and then stopped and let somebody else have "Seven-Up."

[Laughter.]

DR. LIPSEY: At any rate, the U3 overlap index essentially sort of arbitrarily sets a threshold at the mean of the control group. So, by definition, in a symmetrical distribution, we've got 50 percent of the control group above the mean,

right? So, now we've got an intervention effect. In this case, I've shown it as .73—an effect size of .73—.73 standard deviations.

So, what does that mean? Well, what it means is that now 77 percent instead of 50 percent of the students, say, are above the mean of the control group. So, whatever that, the simple average of where you would have been without the intervention, we now have 27 percent more students that are above that point.

Are you seeing that? Isn't that easier to understand than an effect size of .73? It gives you a little better sense of what we're looking at here. And, alternatively, you can see these—we can translate those into percentiles. By definition, the mean of the control group is at the 50th percentile.

What we've talked about is pushing—the intervention has essentially pushed the distributions that the average student in the intervention who's exposed to the intervention now scores at the 77th percentile instead of the 50th

percentile. So, we get a better, a more intuitive picture, I think, of what this effect size might mean.

Just for completeness, another commonly used scheme here is the Rosenthal and Rubin Binomial Effect Size Display. They really work on a similar principle. They set the cutting point here—they're sort of arbitrary, these are just sort of default options—at the grand median or the grand mean.

And we can see in this case, we've got a d of .80, an effect size of .80, and when we do that, it means that below that somewhat arbitrary grand median point, we've got 70 percent of the control distribution, only 30 percent of the treatment distribution, and, conversely, 30 percent of the control distribution is above that point, but 70 percent of the intervention distribution is above that point.

I don't find this particular one somehow as inherently interpretable as at least looking at the mean of the control group because this grand

median of the distribution is going to depend on where the distributions are, and so on.

But it has a cute property, which is that the difference between the proportions on each side, in this case, between 70 percent and 30 percent, difference between .7 and .3 is .4. And .4 happens to be the correlational version of that effect size, and you can see it right there as .4, which, to a rough approximation, is about half of what your Cohen d regular effect size is.

I mean if you really want to impress your friends, I guess, you can kind of do a quick mental calculation here and produce the BESD value very quickly.

But I think generally, if we pay a little bit, if we give a little bit of thought to the context within which we're doing the intervention, that we can set more meaningful thresholds here and translate our findings into proportions that are above—in the respective treatment and control groups that are above some more inherently meaningful threshold.

A good example of this is in the NAEP data, where, through a process many of you are familiar with that involves kind of categorizing the items and using experts saying what somebody should know if they're in the fourth grade or the eighth grade or so on, there are external criteria set for what's, say, a basic reading level, a proficient reading level, and so on.

In this particular example here, we've got basic achievement on—these are the 2005 fourth-grade NAEP scores for free and reduced-price lunch kids. The basic achievement level is a score of 208. So, we can look at, if we had an intervention group at the top and a control group at the—no, it's the other way around, I'm sorry—control group at the top and intervention group at the bottom, we can translate what in this case is an effect size of .20 into the proportion of kids who now read at the basic level.

Before the intervention, without the intervention, 44 percent of them could read at the basic level or above. With the intervention, 52

percent can read at the basic level or above. You could do the same thing at the higher end with the proficient level, which has been set as an external criterion at 238.

Now, without the intervention, 16 percent can read at the proficient level or above. With intervention, that goes up to 21 percent. We pushed five percent of the kids with our intervention over that threshold and now can read at the proficient level when before they couldn't.

You following me okay here? So, a simple thing to do-since we've got the data, we can make the distributions, we can put these cut points anywhere-is to attempt to derive a meaningful cut point, a meaningful threshold within the context of the intervention study itself, and to just produce a translation.

You can produce more than one translation if you want to show this in different ways, but to produce a simple translation of your statistical effects into this somewhat more intuitive form.

And there are a number of options here.

You can be very creative about it. I've talked about, I've just shown you examples of three, but there are other possibilities here.

For instance, we could look at the mean of a norming sample. We often have standardized scores for many of our measures like the Peabody Picture Vocabulary—I'm from Peabody College; I mention that in every talk—has a standard score of 100. That's essentially the mean on the norming distribution so we can look at the standard scores for our treatment group and our intervention group.

I don't know about you, but the groups we work with, most of them, the mean is not 100. The mean is well under 100, okay? But we can still talk about how many in the control condition score at the average of the population—the norming population—and how many more are pushed over that average—over that score of 100—by the effects of the intervention.

We can use as a cut point the mean of some reference group when we're thinking in terms of, say, achievement gaps. We're working with a high-

risk population, say, that a low SES that qualifies for free and reduced-price lunch, well, what's the average score for the students in that context who don't qualify for free and reduced-price lunch?

And how many of—what proportion of the students in the control group score at or above the average of the students you would hope that you would aspire to have them match if you have an effective intervention, and how close do you come to closing that gap? I'm actually going to return to that concept a little bit later.

We can take our arbitrary scales, and we can do a little extra work and try and establish sort of what's a meaningful cross-over point, sort of like clinical significance or clinical levels in the context.

For example, go back to my conflict resolution program and the interpersonal aggression that the students are reporting. We might work with teachers to identify students that are sort of just under the point where they're really a problem for the teacher, and those who are clearly over the

point where they're really a problem for the teacher, and look at their scores on this measure and pool that across teachers and come up with a cut point where, on the one side, students are viewed as problematic in that context, and, on the other side, they aren't.

And then we look at our conflict resolution intervention, and we translate this into proportions, and now we can report how many were in that problem area as perceived by students—by teachers in this context, without the intervention, and how many students now have been moved into the nonproblem area by the effectiveness of the intervention.

There may be other reasonable values in the context that-what's the target value needed for schools to reach their AYP objectives and so on. And basically any identifiable score you can come up with on the basis of something that's meaningful in this context, you can translate your distributions, your arbitrarily scaled scores, for the intervention and control groups, into the

proportions that fall on either side.

I've spent a lot of time on that. It's really a simple concept. You can do it almost all the time, and I think it's one of the easiest things that we might do routinely in order to better communicate what we're finding.

A couple of other—a couple of other things worth mentioning here. Conversion to grade-equivalent scores. You can talk about what the difference between the intervention and the control group is in terms of grade equivalents moving from given grade equivalent to a higher grade equivalent.

Here's an example from one of Bob Slavin's papers on Success for All. We've got grades one through five, the Success for All group, the control group, and in order to better communicate his results on an achievement measure, he basically translated them all into grade equivalents so we can see at the different grades about how much movement there has been in terms of grade-equivalent status.

Grade-equivalent scores do have some characteristics and quirks that one has to keep in mind, however. For one thing, they're provided by the test developer. They're not something you automatically generate for yourself. So, standardized norms tests will often produce the grade-equivalent values for each score.

But if you don't have that, you can't do much with grade equivalents, though it occurs to me for those of us working in context where, say, state achievement scores are available for a whole district or a whole school, we could basically treat that as a norming distribution and come up with grade-equivalent scores for our particular district or the particular school within which we're doing an intervention.

Grade equivalents, as you may know, have a metric that vary from .0 to .9 over a school year, so 4.0 is the grade equivalent for somebody beginning the fourth grade.

A consideration here when we're translating into grade-equivalent scores is that

these are not criterion referenced. They have kind of the sound of a criterion reference value. This is how somebody should score who's performing appropriately up to standard in the fourth grade, but they're empirically derived from the norming distribution.

So depending on what norming distribution was used by the test developer and when, five years ago, 10 years ago, 20 years ago, it's the grade equivalents that were found in that data and may or may not be applicable to the situation you're working with.

Also, the grade equivalent scores in most of these standardized tests include a lot of imputed values. They don't have empirical data for every combination, and particularly for students who are scoring outside the grade range where empirically you'd have to have a student who was in the fourth grade scoring like a sixth grader or in the fourth grade and scoring like a second grader and so on.

So, not all of these grade-equivalent

scores that we get from the test scoring programs are even empirical values.

And we have to think about grade equivalents as a translation of the original scores, not as an alternate metric within which to analyze those scores because they have nonlinear relationships with the actual scores themselves, and we also have to keep that in mind in interpreting them.

So, the equal grade-equivalent differences for children in the younger grades will represent a bigger difference in raw scores than in the older grades, but, conversely, we get more variation between individuals within a grade with the older kids than in the younger kids.

So, if you're looking at two ninth graders, one with a grade-equivalent score of 8.7, another, 8.5, that's a much bigger difference in terms of raw scores than the same difference for third graders, which complicates a little bit the picture, but the general logic I think of translating into something like grade equivalents

that is more easily understood in context is certainly good.

There are other translations that one can think about, percentiles, normal curve equivalents, and variations on that, that we do talk about in this paper that's in the works, but I'm going to turn here to the related question of practical significance.

And the difference here is that for practical significance, you need some kind of external framework, something that gives you criteria for what is meaningful in a practical sense, something else to connect with, so not simple descriptive translation of the statistical results but invoking some external criteria or some external standard that allows you essentially to assess—to evaluate the practical significance.

And actually we see a touch of that when we bring in our own meaningful thresholds into simple dichotomizations of the treatment and control. When we're talking about basic reading levels or proficient reading levels by criteria

that have been defined externally, that actually is getting us closer to practical significance.

Let me show you some—this draws on the work that Howard Bloom and I and a very fine team have done recently. Let me show you some of the frameworks that we think are more interesting and useful for assessing practical significance on achievement scores.

One of the things we might do is to compare the effect size we find in our particular intervention with the distribution of effect sizes that have been found in similar studies with similar interventions and similar outcomes. Is this intervention producing something that's at the low end of what other people have been able to produce in similar situations or, in fact, is this at the very high end? So it's a kind of an actuarial framework here. How do we compare with what's happening elsewhere?

And I want to talk about normative expectations for change. Go back to the pre/post picture we were looking at before and think a

little bit more about that as a framework for practical significance—those policy-relevant performance gaps that we want to enclose between high SES and low SES, or minority and majority, students and actually something I'm not going to talk about further here but needs to be mentioned is an economic framework, the intervention costs relative to some economic assessment of the benefits, which certainly in a policy context is a very direct way of establishing practical significance.

Let me look first at just the idea of comparing the effect sizes found with some kind of effect-size norms from other intervention studies. This is basically what Cohen—to return to the Cohen rules of thumb—this is basically what Cohen was trying to do except he just sort of estimated from his experience what the effect sizes seemed to be in social science intervention research.

I actually, over here on the right, some years ago, with a colleague, we combed through about 350 meta-analyses of psychological/

behavioral/educational interventions with all kinds of different outcomes and all kinds of different situations and plotted out a little distribution of mean effect sizes and divided it into three parts and, by gosh, got about the same numbers Cohen did, and he didn't do anywhere near that much work to kind of dream up his numbers.

[Laughter.]

DR. LIPSEY: So, this is a normative distribution, but the problem, of course, is that it's a normative distribution for a very heterogeneous mix of interventions and outcomes and situations.

The problem is not the idea of making reference to a normative effect-size distribution. It's using the right norms, and this is not a particularly good norm for what we do in education.

Let me show you an example from some—we've been collecting random assignment studies in education, and thanks to IES, there are just a whole lot of those recently, and the number keeps growing.

Did I see Kelly out there in the audience? Yeah. Kelly has done a lot of work pulling this together. So we need to thank him for that. But this database right now has got 124 studies in it, 181 independent groups, and 831 achievement effect sizes.

At least for achievement outcomes, here are some norms that are a little closer to what we might be concerned about in education. Here's one breakdown that differentiates the grade levels. We've got elementary schools, middle schools, and high schools. And most importantly here, differentiates the kind of achievement test.

One category, the two categories of standardized tests, the one we call "broad," these are your really broad band, like your overall reading score, your overall math score from state achievement or CAT9 or any of these standardized tests. The standardized narrow band is typically a subtest, so not the overall reading score, maybe, but the reading comprehension score—the vocabulary score—and typically they will be picked by the

researcher, of course, because they are somewhat better aligned with what the intervention is actually doing.

And then the third category here called the "specialized topic test," these are the more tailored tests. These tend to be the measures that the researchers have cooked up themselves or sometimes teacher-derived measures and not standardized measures at all.

If you look at just the average effect sizes here across for achievement outcomes across elementary, middle, and high schools, you can see that there is some grade-level differences. This is now, this is a whole mix of interventions. I'm not talking about what the interventions are, and they're different interventions at these different levels.

So, I'm not actually going to advocate— none of these numbers are a whole lot better than Cohen's numbers, though they're closer to getting us into the domain of what we might expect with achievement measures. The norms we really want are

going to get more specific to the kind of intervention and the kind of target samples we're working with. And I think it would be a useful thing in the world of education if we had better compilations of these things that we could all look at and have a better idea of what to expect.

And we're currently trying to persuade the IES reviewers to fund, recommend funding for an effort along that line. Notice particularly here the difference between these different levels of achievement tests as we've categorized them, and look particularly at these broad band achievement measures.

We've just lumped together all of the interventions anybody has tried and reported in the literature, and this is only random assignment studies because we wanted effect sizes that we were pretty sure were actually estimating intervention effects and were not confounded with selection bias or other things—.06, .07.

And we don't have enough for high schoolers. The one we have is also very small, but

I didn't report that one. If you take Cohen seriously, that is utterly trivial. But that's what the average of anybody, you know, of all the interventions we have been able to produce when they measure those.

Another way of putting it is these broad-band achievement measures are great big battleships of measures that don't move easily based on any of the kinds of interventions that we're testing. They're probably not particularly sensitive to interventions.

As you go to the more aligned and the more tailored kind of test that the researchers, remember, are selecting because they think they're appropriate for their particular intervention, we're seeing effect sizes that get larger. But by a Cohen-type standard, they're not even medium.

Nonetheless, given the state of the art in education, I have not shown full distributions here, but an interesting thing to do with these distributions is to kind of pull out the 25 percentile level and the 75 percentile level and

narrow it a little bit to target areas that are comparable to the research that you're doing, and then you can compare your effect size.

Is it up around the 75th percentile or more around the median or more around the bottom of what has been found in similar studies with similar interventions in similar populations?

You see the practical significance framework here. It doesn't translate this into dollars or lives saved or anything like that, but it gives a framework for saying is this an intervention which in practical terms is doing as well as any intervention anybody has tested out there or doing a lot worse or doing a whole lot better—is a high end, low end?

But this achievement area is particularly tricky, and for those of you that design studies and do statistical power analysis and so on—and many in this room like me are reviewers on these panels—anyone who comes in with a broad-band achievement measure and says I'm expecting an effect size of .50, has got an uphill argument in a

context like this to make a convincing case that that is realistic.

Here's another cut on that same data, and detailed versions in some ways are more interesting, but the number of data gets thinner. Here, we've made a bit of a pass at not so much differentiating the kind of interventions but at least the level at which the intervention is provided.

So, you've got interventions that are targeted on individual students—kind of one-on-one tutoring-type interventions—and technology, computer interventions, and so on—those that work in small groups outside of classrooms, those like a curriculum at the whole classroom level, and then whole school interventions. You can see quite a bit of variation there.

So, if you're doing a whole school intervention and thinking about what might be a big effect size relative to what anybody has been able to accomplish in changing average achievement levels at the whole school level, an effect size of

.15 relative to what's been found so far would be a knock-your-socks-off effect size. In another context, that would be trivial.

Let me switch then to another possible framework here that is relatively easy for us to apply, and this is expectations for change. This kind of comes back to the pre/post context comparison, and I'll just show you some data here.

This slide kind of overviews what we've done. We've taken norming data from seven of the major standardized achievement measures. So now, these are pretty much the broad-band reading and math standardized achievement measures, and we've looked at what the change is in this norming data. So, this is presumably just population data.

What's the change from year to year, from fall at the beginning of first grade to fall at the beginning of second grade? Or maybe it's spring to spring. Now, I'm not remembering. Maybe it is spring to spring. But, at any rate, it's a 12-month year period.

What is the change on these measures? And

we'll translate that into effect-size units, and then that gives you a framework for thinking about an intervention effect and how much has it added to or accelerated the amount of change that was pretty much normative for that age group?

And if you do that, if you pool this, and actually, there's remarkable similarity in these functions across the different major achievement tests. If you do that for each of these intervals, you get something like this, and notice how this decreases, okay, so the K-to-one transition, one-through-two transition, just from year to year on these achievement measures, we're seeing about a standard deviation gain.

If you have an intervention effect, say, that produces an effect size of .10, that's a 10 percent increase approximately on what—on the underlying change that you would expect normatively as a function not just of the educational experiences of these kids but the whole life circumstances and family. This is just the whole thing; this is the whole developmental change.

If you go up, say, to the ninth or tenth grade where you see for the nine-to-ten transition, you have an effect size for that developmental period of about .2, okay, I don't know—this is an interesting question.

Some of you may know more about this than I do. I don't know if these achievement tests are just no longer measuring change as you get up into high school or if we just have brain-dead high schoolers who aren't changing?

But look at the implications of this. If you do an intervention now with ninth graders and have an effect size of .10 on one of these measures, that's a 50 percent increase on the developmental change—the change over a 12-month period—that one would have expected without that intervention. That's big.

Same effect, you see, but now we're trying to interpret its practical significance with regard to how much it accelerates what is a year's worth of development, and, as you see, it's really quite irregular.

The pattern here—here's actually a plot across all seven of these sets of norming data that we collected with the confidence intervals, and then the maximum and minimum effect sizes in pink and with the little triangles, and you can see actually that there's remarkable comparability across all of these different broad-band achievement measures, and you can also see this is such a regular pattern—this decrease over the years—that you can actually fit it fairly well with a little quadratic function.

So I'll turn now to another possible framework we're thinking about, and this is the sort of the closing the gaps. Now, I'm not advocating any of these as the golden road to assessing practical significance. Different ones will make sense in different circumstances.

There may be still other kinds of frameworks that would make sense in a particular circumstance. My advocacy here really is to think about what is the framework from which one can derive the practical significance of our

intervention findings.

These are examples, but there are other things one might do as well, to think about that framework to try and apply it, and then to try and actually produce some assessment for the users by some standard or another as to what the order of magnitude of the practical significance might be.

Here, we would do a similar thing. Instead of looking for year-to-year change and how a given effect compares to what you would expect a year's worth of development to produce ordinarily, we're going to look at the magnitude of a gap between, say, between demographic groups.

So, here, we've got data that translates into effect sizes for grades four, eight and 12. The first, second, and fourth of these, which are running about .75, .8 of a standard deviation, that's the Black/White gap, and the White/Hispanic, and then, at the end, the free and reduced-price lunch versus not free and reduced-price lunch differential. So, we're just basically looking at the average for those different groups, translating

that into an effect size.

And then the smaller one there is the male/female difference incidentally—the third one there.

So, the framework here in terms of practical significance, if you've got a given intervention effect, say, a .25—these are again broad-band achievement measures—if you've got an intervention effect of .25, what's its practical significance?

Well, that, if you're working with a minority population or you're working with an economically disadvantaged population, that's closing about a third of the gap. That's a framework for—if it's .10, if it's .05, it's not making much difference on that gap.

If you come up with an effect size of 1.0, okay, you better patent and sell that puppy because that solves the policy problem all by itself.

You see what we're doing here? We're looking for something that's meaningful in policy terms or developmental terms or best anyone has

been able to do with an intervention terms, and comparing what we're finding in our study in that framework to try to get some idea of whether or not we've got something that has practical significance and how much practical significance it might be.

Here's another kind of quirky performance gap picture that puts in perspective some of our achievement effect sizes. This is a variation of something that Tom Kane and then Larry Hedges have written about, but what's the difference in achievement measures between average schools and low-performing schools by the usual definitions?

And, of course, there are different students in those so you kind of do some regression and statistical manipulations to try and control for the different student backgrounds so the statistical image you're trying to produce here is that we've got equivalent students in a high-performing school and a low-performing school, and what does that difference in experiencing high-performing—actually, an average-performing and low-performing school—what does that translate into in

effect-size terms on achievement measures, and then we might use that as a basis for comparing.

If we imagined that we were working with students in a low-performing school, and we had an intervention with a certain effect, how much closer would that push them to performing like the students—equivalent students—who have the benefits of an average-performing school?

Here's some data that comes from the MDRC archives that Howard Bloom and his colleagues produced for two different districts, and you can see that those numbers range in effect-size terms from across the grades from about .15 to up around .30.

So, in that context, if you had an intervention that would push a whole school average up by about .2 standard deviations, the practical significance of that, I think you would have to agree, would be pretty huge. That would be like taking a low-performing school and having it generate achievement scores that are average for the schools that are viewed as adequately

performing.

Well, this is not the whole story. These are just examples, but just to kind of reprise a moment, and then I will stop, and I think we have time for questions, don't we? Good.

So, my main points here that I hope are obvious enough—the native statistical form that we get naturally out of our intervention studies is, by and large, uninterpretable in any intuitively meaningful form, largely for us researchers, in many cases, I guess, which is why we glom on to p-values with such vigor, or maybe that's just because we can get them published.

But certainly for consumers, practitioners, and policymakers, who in these arbitrary scaled units are going to have trouble making any sense out of what we found, and the fact that something is statistically significant is no assurance whatsoever that that effect has any practical significance or is even nontrivial.

So, translating those native statistical forms into something more descriptive at least

gives us and any consumer a better idea of what it might mean and what it actually is, and what I've tried to illustrate with the various examples is that's really not all that hard to do if we think about doing it routinely, and we think about what makes sense with our particular data and our particular circumstance.

There are a number of ways we can—just in another line or two after the basic statistical results or in the discussion section or the executive summary of the report or so on—give a more descriptive or comprehensible picture of what the nature and magnitude of that effect was.

And then to try and move a step further into practical significance, we need to bring in some practical framework or practical criteria, and that's going to make things more difficult. There are a number of those that might be appropriate. All of them, for the most part, are going to require that we derive or develop or borrow some kind of information or some kind of criteria that are meaningful in that context, so it's not a

simple translation.

So, that is certainly more difficult, but there are a number of relatively straightforward ways–and I'm sure a creative group of researchers could come up with many more than have been developed in the literature so far—that do not require a huge effort. And with many of the kind of outcomes we work with in the world of education, there's a lot of data and a lot of sources out there by which you can derive some of these frameworks.

So, it wouldn't be that difficult for us to routinely pick an appropriate framework and actually address the question of assessing practical significance for each of our intervention studies.

And, I think if we at least translate and make some attempts to assess practical significance, that we will better understand our own, the implications of our own research, and we will certainly better communicate with those that we are hoping draw on the body of educational

research in the world of practice and policy.

And that's about as far as I can get this morning.

[Applause.]

DR. LIPSEY: I'm happy to take some questions. I guess they're recording this, so they've asked that all the questions go on the microphone, and I guess I need to answer them on the microphone, too.

DR. WOLF: Okay. Patrick Wolf, the University of Arkansas.

Mark, what do you think about for longitudinal RCTs over multiple years the translation of effect sizes into additional months of learning based on the actual sort of demonstrative magnitude of learning across years for the control groups in the study?

DR. LIPSEY: Yeah, yeah. Great question. We are working on that. Michael, would you like to answer this question?

[Laughter.]

DR. LIPSEY: To frame it a little

differently in terms of translating, you know, say we've got one of these multilevel things with growth curves, and we're actually looking at our intervention effects in terms of differential rates of change over time, what are the ways to translate that?

And you suggested one. And we're actually pulling together what little literature there is on that and trying to do some creative thinking. I think that's a particularly challenging area, but there is no—you certainly—I think it's obvious that the regression coefficient that tells you what the difference is between the slopes—the average slope for the treatment group and the control group by itself—is not going to communicate very well to anybody.

So going back to the—if the original metric is meaningful—going back to the original metric and at least talking about rate of change. But often the original metric you see is not meaningful, so where do you go from there to get into a rate of change framework that has a little

more intuitive meaning?

If you have any ideas about that, I'd be interested to hear them, but I think that's a particularly challenging area for these translations.

AUDIENCE PARTICIPANT: I commend you for trying to make measures of strength of effect practical, understandable, but almost every educational intervention has multiple outcomes, and I still worry constantly about how do you convey the total impact of even the most narrow intervention, not to mention broader things like whole school reform or, you know, suppose you have an intervention that increases—a classroom intervention that increases every achievement score by .2 of a standard deviation? Certainly, that's a far more powerful intervention than any one effect size would convey.

DR. LIPSEY: Yeah. That's a really good point. Everything I've talked about, we could attempt to apply outcome by outcome even in a multiple outcome study, and, of course, not every

study measures all the outcomes that might be affected by the intervention, so we don't have much comparability across studies.

That's an interesting problem, though. How would you characterize descriptively and in terms of practical significance effects on sort of a front of effects that might be not just different achievement areas, but some interventions might produce significant changes on achievement and on classroom behavior, for example? And obviously, that has more significance than an intervention that—practical significance—than an intervention that only affects one of those and not the other.

I don't have any good insights on that, but I think that would be an interesting area to explore, is how we might capture those cross-outcome impacts in a more descriptive way.

Greg.

DR. DUNCAN: Hi. Greg Duncan, University of California-Irvine.

You preempted my question on costs and benefits; although, in response to the previous

question, I would suggest that assigning dollar values to a range of benefits is one way of adding up different kind of effect sizes.

DR. LIPSEY: Yeah. Good point.

DR. DUNCAN: So—

DR. LIPSEY: If you can do that.

DR. DUNCAN: Right. If you can do that.

DR. LIPSEY: Yeah.

DR. DUNCAN: You know, for things like grade failure, for things like special education placement, if you can get impacts on those rates, it's pretty straightforward to value what a year of education costs, what the incremental costs associated with special education might be, to translate at least part of the effects into dollars.

DR. LIPSEY: Yeah. If I could just comment a little bit on the costs. That framework bothers me a little bit because so many of the educational outcomes are difficult to value and require a lot of assumptions.

I think that the safe ground for me is

where you can translate it back into systems costs. If a kid is retained in grade, you know, you've got to provide the teachers and the classrooms and the resources to run that kid through the fourth grade a second time or so on, and there's an immediate cost to the system and other things like that.

That, on the benefits side, you know, what the savings are of changing grade retention or special ed placements or something, that seems very straightforward. When we start getting into other domains, as you well know, you know, what the economists start looking at is lifetime earnings and other things like that, which is quite common, and that does provide a common metric.

I guess I'm one of the people that are a little uncomfortable with the idea that the purpose of education is to increase lifetime earnings.

DR. DUNCAN: What else is there?

[Laughter.]

DR. DUNCAN: So, that actually wasn't my question.

DR. LIPSEY: That wasn't your question.

DR. DUNCAN: Right. My question was about confidence intervals.

DR. LIPSEY: Aah.

DR. DUNCAN: Right. And how do you convey some sense of the reliability of the estimates and, in particular, in the context where, say, you've estimated a number of effect sizes, they're all generally pointing in a positive direction. Some are clearly significant; some aren't. How do you handle kind of the marginal borderline insignificant effect sizes? It's clearly not zero. So have at it.

DR. LIPSEY: Well, I think you and I probably share a lot of common attitudes about that. The state of practice around this kind of crude, brute force dichotomy between what's, you know, .051 and what's .4049 on statistical significance is really pretty silly, and I don't know if we're moving to something more sensible or not. That's very well entrenched. The confidence interval part, you know, I certainly think confidence intervals are much better than p-values.

In the context, though, of what I've said here, a confidence interval doesn't communicate any better than a p-value in terms of—descriptively—in terms of the order of magnitude. I think what it would help you do, any of these descriptive pictures we might paint, if we really wanted to show the range of uncertainty, at least as we could estimate it statistically, you might do that for the lower range in the confidence interval and the higher range.

So, if we're looking at our estimate of the proportion of kids that are pushed over a meaningful performance threshold, for instance, we could actually do that across the confidence interval range and probably give a fuller picture of what we found, and I think that might make sense.

If I've got enough, if I've got enough of those effects with p-values, you know what I'm going to do. I'm going to put them into a meta-analytic framework and try and make an aggregate estimate for those, and there can be technical

complications with that as well.

But, you know, the bigger picture here is that we need to be looking at the empirical results for the magnitude—the nature and magnitude of the effects that are being produced—not be quite so obsessed with whether or not they're over or under the margin, some margin, arbitrary level of statistical significance.

But I don't know how to get there, from here to there, Greg. Maybe you have some ideas.

DR. CONAWAY: Hi. My name is Carrie Conaway, and I'm the Research and Planning Director for the Massachusetts Department of Elementary and Secondary Education. So, I'm often a receiver of research along these lines.

We recently commissioned a study with the Boston Foundation about the impact of charter schools relative to traditional schools in Boston where the researchers tried to use the very approaches that you've described here today, and one of the comparisons kind of bugged me, and I'm curious whether my intuition about it is right or

not.

So, they were using the approach of trying to see how much impact did the charter schools have relative to the achievement gap. And they, sort of, they took a generalized view of the achievement gap. They sort of said, in general, the achievement gap is about a standard deviation. The impact—we actually had quite large impacts. These were on the state standardized—the MCAS, our state tests.

In middle schools, the charter schools were getting on average a half a standard deviation per year improvement in student performance. And so, they said the charter schools, therefore, are closing the achievement gap by, you know, half of the total gap is closed in just one year, and it seemed to me that that would only be true if like the poor kids—if it was only the poor kids or the Black kids—that were receiving the treatment of the charter school and only the advantaged students were in the traditional school.

And I'm curious if that's the right way to think about it or not?

DR. LIPSEY: Yeah. Interesting point. Of course, in the background, before we took that too seriously, we'd want to make sure that the claims about the gaps were based on the same kind of measure and the same kind of context, for instance. They could be quite different in your context.

But the idea of practical significance here, and I think you're quite right in terms of appraising the claim that that effect is closing the gap. I think the correct wording on that is that the order of magnitude of that effect would be equivalent to this much of a reduction in the gap.

That's different than producing that much of a reduction because the "equivalent to," as I think you correctly figured out, assumes that the performance of the normative group stays constant. If you have an intervention, you know, the rising tide is floating all the boats.

You may well just push everybody up, and the gap remains, you know, so the fuller picture of that is going to—for that particular situation—is going to have to look at the effects on both groups

and the gap itself. Am I making sense?

DR. CONAWAY: Yeah.

DR. LIPSEY: So, I think you're quite right in terms of the data and whether or not it shows that the gap in that situation was actually closed.

The way I was talking about this is we would take an intervention effect of that sort, and we'd say, gosh, is half a standard deviation—is that trivial? Is that big? What do we compare that with? What kind of yardstick do we use?

Well, one yardstick is that the SES gap—the Black/White gap—is three-quarters of a standard deviation or a little more so relative to that. This looks like a pretty big effect. That's not the same thing as saying that that intervention has closed that gap.

DR. CONAWAY: Okay. Thank you.

DR. LIPSEY: Does that make sense?

DR. CONAWAY: Yeah. I feel better now; thank you.

[Laughter.]

DR. STARNER: Hello. Thad Starner from

Georgia Tech.

I'd like to address the previous question that I thought was really great about lots of small effects on many different outcomes.

DR. LIPSEY: Yeah.

DR. STARNER: I've been doing a lot of work with fMRI studies recently where they have many, many, many different voxels that are lighting up in the brain. They're trying to see some effect across many different voxels, say of thousands and thousands of these things. They have positive Bonferroni correction, Benjamini-Hochberg, and all that sort of stuff.

What they've come up with, something called permutation thresholding, and I think that for that problem, permutation thresholding might be the right way to get all these small effects and show they are significant, and it should be quite stunning. If you really do have that effect, it should show up really well.

So, maybe we could talk about that afterwards, but—are you familiar with permutation

thresholding?

DR. LIPSEY: No, no. I don't know that. That sounds interesting.

DR. STARNER: Okay. It's really surprising and powerful. I can talk to you afterwards about it. Thank you.

DR. LIPSEY: Yeah. One thing I would just say, in general, I don't know that it applies to that technique, but one complication when we're trying to kind of summarize across multiple outcomes on multiple—well, multiple outcomes for a given intervention is that those outcomes are not necessarily independent so one of the things we're going to have to pay attention to is what the interrelationships and correlations are.

You can basically have one effect that you measure in 14 different ways, and that's not 14 different outcomes. That's one outcome measured in 14 different ways. So, kind of figuring out what are truly independent separable effects that need to be taken account when you're looking at the whole front and what are really just highly

correlated variations on the same effect is going to be a little bit challenging, I think, in any context where we try and do that.

DR. STARNER: I think this would probably take care of that, but it would take too long to explain the method.

AUDIENCE PARTICIPANT: Just to follow up on that, I hope you will solve this problem for us of multiple outcomes.

[Laughter.]

AUDIENCE PARTICIPANT: Because IES is sitting here trying to compare different interventions, each of which has that problem. It's like our measure of strength of effect isn't up to our research designs and need for information in this sense, but we did play around with multivariate effect sizes, you know, multivariate distance measures and so on.

It doesn't hack it because if you get a long list of .2s and the variables are highly correlated, your multivariate effect size is .25, and then you're nowhere.

DR. LIPSEY: Yeah, and it still doesn't address the interesting question of the breadth of the impact, you know, with a given multivariate effect size.

You know, it might be still across a fairly narrow range of outcomes or very broad range of quite different outcomes that it would have different significance.

AUDIENCE PARTICIPANT: Exactly.

DR. LIPSEY: I think the general approach—and I confess this is not something I've thought a whole lot about or worked on—but I think what Greg suggested earlier, translate it all into dollars, you have to worry about what are the independent effects and kind of add that up, doesn't necessarily mean dollars are the right—I shouldn't say right, but may not be the appropriate metric.

But I think the way that's got to go is to find some kind of common meaningful metric in which one can identify what are the independent effects and translate into that metric and try and talk kind of cumulatively about the net effects.

But what those metrics might be in different situations, I think, is going to be a bit of a challenge. That would be a good thing, too—all you doctoral students in the room, there's a good dissertation topic.

AUDIENCE PARTICIPANT: Or the speaker.

[Laughter.]

DR. LIPSEY: I'll add it to the list. Okay. It appears that we're done here. Thank you all very much.

[Applause.]

[Whereupon, at 11:41 a.m., the panel session was concluded.]