Institute of Education Sciences

Fourth Annual IES Research Conference
Concurrent Panel Session

"The Problem of False Discoveries: How to Balance Objectives"

Monday
June 8, 2009

Marriott Wardman Park Hotel
Thurgood Marshall North
2660 Woodley Road NW
Washington, DC 20008

Contents

Moderator:

Amy Feldman Farb
National Center for Education Evaluation
and Regional Assistance (NCEE) 3

Presenters:

Peter Schochet
Mathematica Policy Research, Inc. 5

John Deke
Mathematica Policy Research, Inc. 15

Discussants:

David Judkins, Westat 35

Jeffrey Smith, University of Michigan 40

Q&A 47

Proceedings

DR. FELDMAN FARB: Hi. Good morning. I'm Amy Feldman Farb with the National Center for Education Evaluation and Regional Assistance at IES.

And I'm sorry, I know everyone is still trying to make it in here and sit down, but we've lost about 25 minutes, and we want to keep this to schedule, so we're going to have to really try to speed things up this morning.

I have a list of podium notes here. Probably I was supposed to read it to you, but they are the same ones you heard downstairs. So just try to remember to silence your electronic devices. This session is being recorded, so if you have any questions, you're going to have to come up to the mics to ask your question, and you're going to need to identify yourselves before you ask your question.

I'll try to be very brief. NCEE is responsible for advancing rigorous and scientifically valid education evaluations and for translating findings for decisionmakers and practitioners. In 2006, the center awarded a contract to Mathematica Policy Research to provide analytic and technical support for advancing education evaluations.

Among the objectives of this contract are to provide appraisals of the methodological issues that confront ED evaluations and to propose and conduct investigations to ensure that IES is investing in rigorous and cost-effective study designs.

These studies result in our Technical Methods Reports Series. The reports go through rigorous peer review in the IES Science Office, and

once they're approved, we post them online on our website. I can tell you later how to get there if you have a question.

To date, the contract has produced four published Technical Methods Reports on topics including: *Multiple Comparisons*; *Regression Discontinuity Design*; *The Late Pretest Problem*; and *Cluster Estimation Approaches*.

We have three more reports currently in IES review, including: *Handling Missing Data* and *Power to Perform Mediator Analyses*. And we have over ten ongoing studies that should produce reports in the next 2 years.

This morning's session is based on our first Technical Methods Paper: *Guidelines for Multiple Testing and Impact Evaluations*, by Peter Schochet.

This report presents guidelines for researchers that address the multiple comparisons problem in impact evaluations in the ED area. The multiple comparisons problem occurs due to the large number of hypothesis tests that are typically conducted in an evaluation study, which can lead to spurious, significant impact findings. That is the problem.

Although this was the topic of our first published report, NCEE and IES continue to revisit this topic as we strive for the most appropriate and responsible methods of reporting study findings.

The presenters today will first summarize the multiple comparisons approach from the report and then move into a discussion of the more technical issues about the proper correction procedures for between-domain analyses.

We have two discussants today, and then we'll try to squeeze in as many questions and answers as we can. We thank you for being here this morning. Let's get started.

DR. SCHOCHET: Hello. I'm Peter Schochet from Mathematica, and I'm going to be presenting with John Deke, who is also from Mathematica.

I'm going to talk very fast because we've lost 30 minutes. So I want to talk about the multiple comparisons problem and where we're at. This is a problem that IES has been struggling with for the last 2 years or so. It's a very complex and difficult problem. So I'm just going to talk about what the guidelines are and some applications and where we're at right at this moment.

What is the problem? As we all know, when we do multiple impact evaluations, we do multiple hypothesis tests. We look at multiple hypotheses across multiple outcomes; across subgroups; males/females; by age; and even by multiple treatment groups.

Now, that's all fine and good. When you do standard testing methods—though—when you do a test at the five percent level for each individual test, you have a five percent chance of finding a spurious impact estimate.

But when you look at them together, there's a much higher probability of finding a spurious impact estimate. So this clearly could lead to incorrect policy conclusions.

So what I'm going to talk about today is a little bit more about the multiple comparisons problem; what the exact testing guidelines are that

IES has adopted; examples of these guidelines; and how they're being used by the regional economic labs; the educational labs; and the RELs, who are conducting 25 randomized control trials. And then we're going to conclude.

John is going to talk about new guidance on statistical methods for between-domain analyses. I'll explain what that is in just a moment.

We assume a classical hypothesis-testing framework where you do standard t-tests for each of your hypothesis tests, and the test is for the jth one as you're testing at the impact of zero.

So you do your hypothesis test. You do your t-test, and you reject the null hypothesis if the p-value of the t-test is less than alpha, which is usually set at .05. What that means is that the chance of finding a spurious impact is five percent when you consider each test alone.

However, if you consider these tests together, and there are no true impacts—there's no impact anywhere—there's a much higher probability of finding a significant t-test. So this first column shows the number of tests, and the second column shows the probability that at least one t-test is statistically significant. Remember, there are no impacts in this scenario.

So if you have one test, it's a five percent error rate. If you have five tests, there's a 23 percent chance of finding statistically significant impact. If you go up to 50 tests, there's almost a certain chance that you're going to find a statistically significant impact.

What's the problem with this? It can lead to publishing bias. As we all know, we're all looking for positive findings. You might pull out that one finding for that one subgroup, for that one outcome on page 393, and put

that in your executive summary when it's likely to be spurious. There tends to be a focus on the "stars" or the significant impact findings.

There is a great amount of literature on different adjustment procedures for adjusting for multiple testing, and basically what these procedures do is lower the alpha levels for individual tests, and so these methods control the combined error rate or the family-wise error rate so that the probability of finding at least one spurious significant impact is at five percent or less.

There are many available methods. The most common is the Bonferroni, where instead of comparing your p-values to a .05 level, you compare it to a .05 divided by the number of tests. So, for example, if you have ten tests, you're comparing your p-value to .005 rather than .05.

Here are some other ones that are well known in the literature. I highlight the resampling methods, which we'll talk about later. Those have nice properties that they can adjust for correlations across the test statistics.

Those methods are all fine and good, but the downside is that by reducing Type I error, you're increasing Type II error. Another way of saying it is that these methods reduce statistical power—which has the chance of finding real effects.

Here are some simulations which show how serious the problem can be. The first column shows a number of tests, and the second two columns show statistical power if you don't do anything. That's unadjusted, or if you do the Bonferroni method. So you can see if you don't do anything, you have an 80 percent power. That's usually what we power our studies to be able to

detect.

If you use the Bonferroni, you can see with five tests, power goes down to .59. If you have 20 tests, power goes down to .41. So again, a .41 means that the chances of finding a real effect are only 40 percent. So there's this dastardly tradeoff between Type I and Type II errors.

So, with that in mind, we developed with a distinguished panel some basic testing guidelines that balance these Type I and Type II errors. What are these guidelines? Well, first, the problem should be addressed by structuring the data. The structure will depend on what the key research questions are, what the previous evidence is as far as intervention effects, and the nature of the outcome measures, as well as the conceptual theory underlying the intervention.

Importantly, the adjustment should not be conducted blindly across all the contrasts. In other words, you shouldn't just line up all 6,000 hypothesis tests and use one of these adjustment procedures, because you lose incredible amounts of power, and you'll never find anything that significant. So the structuring of the data is critical to this plan. I'll give more detail about what structuring really means.

Another critical issue that's stressed in the medical literature is that the protocols—the plan—must be specified up front, and the purpose of doing that is to avoid fishing for findings at the other end. You need to specify up front what the plans are. The study protocols should specify what the data structure is; what the confirmatory and exploratory analyses are—I'll describe what those are in a minute—as well as what the testing strategy is.

But the main point is that it needs to be specified up front.

In terms of structuring the data, the key element is to delineate separate outcome domains; to put your outcomes into separate domains. How you do that? Well, first of all, it should be based on some conceptual framework, and each domain should represent sort of key cluster of constructs.

You might consider these domains as the elements; as items that are likely to measure the same underlying trait. That is they should have high correlations. For example, you might want to put math test scores in one domain, reading test scores in another, teacher practices in another; student behavior measures, because each of these domains may be measuring the same sort of underlying trait.

The testing strategy should have both confirmatory and exploratory components. The confirmatory component is really what's new here. The confirmatory component is supposed to address the central study hypotheses. It's used to make overall decisions about the programs. It's really the main finding.

If you had to tell a congressman your results, what would you tell him in 15 seconds or less? That's your confirmatory component. You need the statistical rigor there, so you have to adjust for multiple comparisons in the confirmatory component.

The exploratory component is sort of everything else. It's to identify impacts or relationships for future study. No matter how plausible or how reasonable they seem, the findings should be regarded as preliminary. So

you're sort of hedging your bets on the confirmatory component, although the exploratory component is important for identifying new, unexpected findings that could be tested in the future.

The focus of the confirmatory analysis in randomized control trials has to be on experimental impacts; that is treatment control difference. And IES has mandated that the focus needs to be on child outcome measures such as test scores.

It's also okay for the confirmatory analysis to focus on targeted subgroups if the intervention is aimed at particular types of students, for example, ELL—English language learner—students.

Not all experimental impacts could be confirmatory. Some can be exploratory. Subgroup impacts are likely to be exploratory in most cases, as well as secondary child outcomes, and teacher outcomes could be exploratory. So again the confirmatory analysis should have a small number of key outcomes for the study.

What's the testing strategy? Well, the confirmatory analysis has two parts. As I described before, the basic plan delineates separate outcome domains. So you might consider testing each domain separately as well as conducting a between-domain analysis.

So regarding domain-specific analysis the idea is to test impacts for outcomes as a group. The outcomes within a domain, are supposed to be measuring the same underlying construct. So you gain more power and you help reduce the multiple comparisons problem by testing these outcomes as a group.

So the key recommendation is to create a composite domain outcome. In other words, create a single domain outcome that is a weighted average of standardized outcomes within the domain. There are various ways that you can form a composite. You can weight each outcome the same. You could use expert judgment. Probably the best method is to use predictive validity. In other words, correlate the specific outcomes with longer-term outcomes and use those sorts of regression weights.

You could do a factor analysis, and you could do a multi—a MANOVA model—but we don't recommend that for some technical reasons which I don't have time to discuss.

So, the idea then is very simple. You have this composite outcome and you just conduct a t-test on the composite. This is a really nice way of reducing the multiple comparisons problem to just one dimension. So it has that nice feature about it.

Now, IES has also mandated in general that a between-domain analysis is necessary. That is, you need to test across domains to get summative evidence as to the effectiveness of an intervention.

The testing strategy for this between-domain analysis will depend on what your main research questions are. If your main research question, for example, is that all impacts may need to be significant, in that case, you don't need to do any adjustments. You just need to ensure that each of the domains is significant.

The more common question is whether impacts are significant in any domain. I'm going to call this intervention a success if at least one of

these domain impacts are significant. Then you need to do multiple comparisons adjustments and we're going to talk about that later. That's what John is going to talk about.

Before turning to that, though, I'd like to just quickly talk about the application of the guidelines by the Regional Educational Laboratories. It's really remarkable that the IES has funded 25 randomized control trials across the country; and these randomized control trials are specific to each region, with each region having picked the trials they thought pertained to their particular issues.

There are 25 of these trials that are currently underway. The basic features of most include having a single treatment and control group; they're testing a wide range of interventions. There is teacher professional development in math, reading, economics, and science. There is school restructuring. There are student assessment interventions. There's a wide range of those.

They typically cover grades k through 8. Most of them are for a single year, but not all of them. Some of them are longitudinal. They collect data over 2 or even 3 years. And all of them are effectively collecting data on teachers and students.

We've been involved in reviewing all of their materials, and each RCT, they've have to provide a detailed analysis plan to IES. And they needed to provide information on what the confirmatory research questions are; what the domains are and the outcome measures within each domain are; their testing strategy for both the within and the between-domain analyses;

what their study samples are; and whether their confirmatory analyses have enough statistical power.

They can't specify, for example, ten separate domains because they won't have enough statistical power. So these analysis plans have had to be structured and focused, and there had to be an element of discipline that is specified up-front rather than at the other end to avoid any kind of multiple comparison problems or fishing for findings.

What are key features of these, of the confirmatory, of the domains? Almost all of them specify student academic achievement in all these 25 RCTs; some domains pertain to behavioral outcomes; some pertain to specific time periods for longitudinal studies; and some pertain to specific subgroups that are being targeted by the intervention, such as for ELL students.

It's remarkable that these guidelines have been applied in the field and have actually been successful in structuring the analyses up-front. So every RCT—we're still in discussions with some of them—but they have specified structured research questions. Most have fewer than three domains. Some have only one domain. They've reduced their confirmatory questions to just one domain, and most of them only have a few outcome measures. So these analyses are very focused and targeted.

The main between-domain question for sort of assessing summative evidence across the domains is, are there positive impacts in any domain? In order to answer that question, one needs to apply some kind of multiple comparisons correction across the domains, and that's what John is

going to talk about, so I'm going to turn it over to him right now.

DR. DEKE: Thanks, Peter.

I'm going to be talking about four different methods that can be used for cross-domain adjustments. I'm first going to talk about what these methods are all trying to achieve. I'm then going to describe those methods, and finally I'm going to show an assessment of the performance of those methods based on some simple simulations that we ran.

First, what are these methods trying to achieve? Well, I think Peter has talked about this, but I'll just restate it here. The goal here is to avoid making a mistake or to control the probability of making a mistake. The fundamental mistake that we want to avoid is the mistake of implementing an ineffective intervention, and so we want to control the probability of making that mistake. And we think that in terms of the types of error rates that have been defined, the one that is most closely aligned with that objective is what they call the family-wise error rate.

The family-wise error rate is the probability of finding at least one significant impact when, in fact, there are no impacts—when there are no *true* impacts—because it is in that circumstance where you would implement an intervention mistakenly.

There is another concept of an error rate that is out there called the false discovery rate, which I'm not going to talk too much about. It's basically where you have a large number of statistically significant findings, and you're trying to control the proportion of those findings that are not true. This is an interesting rate to think about when you're looking at a

preponderance of evidence kind of circumstance, but we think that the family-wise error rate is the right thing to focus on for our purposes, and so the methods that I'm going to be talking about today are all targeting that family-wise error rate.

So what are the four methods? The first method is Sidak, or Sidak. I'm not actually sure of the correct pronunciation. This provides an exact adjustment when test statistics are independent.

The next one is the Bonferroni, which provides an approximate adjustment when tests are independent. The neat thing about the Bonferroni is that it is such a simple adjustment that is still very close to being right when test statistics are independent, and it's one of those sort of great things of nature that something simple really works pretty well.

The third method I'm calling "generalized Tukey," even though I'm not sure that it has a formal name. The advantage of this approach is that it adjusts for correlations between test statistics, when the test statistics are not independent. I'll explain why I'm giving it this name later.

The fourth approach is resampling, and this also accounts for correlations between test statistics, but it's also robust to other deviations from standard assumptions. So you don't have to assume multivariate t-statistics in order to use the resampling approach.

So in looking at these four methods, we have a couple of research questions here. First, what are they and how do they work? And then, second, are the more complex methods, which is this generalized Tukey and the resampling, worth the extra effort? If there was no effort involved with using

these, then obviously you would want to use them because they do provide a benefit, but there is a cost, which is there aren't simple turnkey seamless solutions in existing software packages to use these other methods, and so we want to try to figure out what are the circumstances where the extra effort is worth it?

The basic set-up for thinking about these different methods is that in each case, we're trying to adjust for comparisons across N domains where we have a single composite from each domain. We want to test whether any of these domain composites, if there's a significant effect on any of the composites. We're trying to control that family-wise error rate at the five percent level.

And then finally, we're going to sort of think about all of these methods in terms of an adjustment to the alpha level for the individual tests. That is literally how a method like Bonferroni works. You adjust the alpha for your individual hypothesis test in order to control the overall FWER. The more complicated methods don't exactly work that way, but they can be sort of presented in that way, which then makes them comparable to the other two methods.

I'm going to march through these methods. The first one, the Sidak, exactly controls the FWER when the tests are independent, and so what we've got here is a restatement, which we realized this morning was actually slightly misstated, but I didn't have a chance to fix it on this version of the slides.

The FWER is the probability of incorrectly rejecting at least one

of the tests, which could be restated as one minus the probability of rejecting none of the tests under the null hypothesis, or in a not quite right way to say it, accepting the null hypothesis in all cases.

And then there's an easy translation from that restatement of the FWER into the next line here, which is when these things are independent, you can express this as 1 minus the quantity 1 minus alpha to the N.

Now why does that work? Well, if you think about an individual test, and you're testing it at a five percent level, then at 1 minus alpha, it would be .95, and .95 is the probability of not rejecting the null hypothesis when the null hypothesis is true.

And so what's the probability of not rejecting the first, the second and the third test? Well, we know from interest statistics that the probability of event A and event B and event C all occurring at the same time is the multiplication of those three things together if they're independent.

So this is a nice restatement of the FWER, and then the Sidak is going to control the FWER by choosing the individual alphas so that the FWER is the value you want it to be.

So we just set .05 equal to 1 minus the quantity, 1 minus alpha to the N, where N is known. You've got a single equation and a single unknown variable and so anyone in high school algebra—well, not anyone—but—

[Laughter.]

DR. DEKE: —people who do well in high school algebra would be able to solve for that value.

And so we've got an example here. If N is equal to 3, if you solve

for alpha, you get .017, and if somebody is checking my calculator, you'll find that I've actually rounded it. There is a whole bunch of other decimals.

And then we can state what the adjustment is that you would have to conduct to the individual test—this factor adjustment—in order to apply this, and it's 2.949. So that's the explanation of Sidak. It provides an exact control for the FWER when the test statistics are independent.

Now, the neat thing about Bonferroni is that, although Sidak is not too hard—an accomplished high-schooler could do it—it's even easier to do Bonferroni because Bonferroni is just dividing the alpha by the number of tests. And what's really pretty remarkable is that even though this is a simple approximation, it's a pretty good approximation. So the Bonferroni adjustment factor is the same as N, and then you can see Sidak in the middle, and you can see, even as we get to larger and larger numbers of tests, Bonferroni is pretty close to Sidak. So it's a good quick and dirty method when the tests are independent.

When the tests are not independent, though, both of these methods are too conservative. And there are two situations where you could have correlations between test statistics. And as I was preparing my slides this morning, I realized I could spend 15 minutes just talking about these two sub-bullets and explaining why this is, but I don't think that's probably what most people are going to be interested in. So I will say 'take my word for it,' and in the Q&A, if you'd like more of an explanation, I could give it.

But there are two scenarios where your test statistics could be correlated. One is where the outcomes are correlated themselves, but even if

the outcomes are not correlated, if you have heterogeneous treatment effects, that could also lead to correlations among your test statistics.

So if you don't take into account these correlations, and you assume they're uncorrelated, you actually lose statistical power because you use a bigger adjustment than you need to use—and the reason that taking into account correlations helps you, you can kind of think about it if you imagine the extreme case.

If you have two outcomes, a math score and a reading score, and you want to adjust for those two things; if they're 100 percent correlated—if every kid's math score is identical to their reading score—then you essentially have one measure. You don't really have a multiple comparison problem. You really just have a single impact.

And so lesser degrees of extreme correlation also give you power if you take it into account.

Another thing that's worth pointing out here is it doesn't matter if the correlation is positive or negative. Either one of them, taking it into account, will give you more power.

All right. So the two methods that we have here that take into account correlations among test statistics are this generalized Tukey and the resampling.

And basically, here's how they both work: If you think of p-i as the p-value from Test i out of N tests, what these methods are essentially doing is they are expressing the family-wise error rate as the probability that the smallest of these p-values is less than or equal to five percent, conditional

on the null hypothesis of no impacts being true.

So this is just a restatement of the FWER. And you can see that, because what does it mean to declare at least one test to be statistically significant? Well, that means that at least one of those tests has to have a p-value that's less than .05 on that nominal p-value. So if you can control the probability that the smallest of those p-values is less than or equal to .05, then you will have controlled the FWER.

Another way to state that is if you can control the probability that the biggest t-stat is above a critical value, then you'll also be controlling the FWER. And that's really what the connection is with Tukey, because Tukey, and also Dunnett back in the '50s, had developed testing procedures, testing adjustment procedures, in the context of multiple treatment groups, and when you do all paralyzed comparisons among multiple treatment groups, those comparisons are correlated, because they involve some of the same treatment groups.

And so they calculated what the distribution was of the maximum t-statistic, given the correlation structure that arises from an experiment of that sort, and they published a bunch of tables in a textbook, and people could go to those tables and look up the appropriate critical values.

This generalized Tukey, as I'm describing it, is essentially exactly the same method except instead of using a specific set of correlations that they use that corresponded to the case they were interested in, this can use any correlations from whatever your particular situation is. So you just need a correlation matrix, and you can apply this method.

The reason they didn't do it back then is not because this idea didn't occur to them; it was because they didn't have the computer power, because you have to numerically integrate a multivariate t- distribution. And so you know, you could call it the Intel method because they made the microprocessors that made this possible.

But generalized Tukey assumes that the tests have a multivariate t-distribution which is pretty standard. I mean, we usually assume that our test statistics have a t-distribution. We wouldn't normally consider that to be a harsh assumption. And then if you have a set of—when I say known correlations, I don't mean that you were born knowing them; I mean that you can estimate them—so you've had data and you've estimated those correlations.

There is a package in the statistical programming language/environment R, called MULTCOMP, which will implement this method. And I will explain how this works. So I guess I should say that R, for those of you who aren't familiar with it, is an open-source programming language/statistical environment. If you do a Google of the word Cran, C-R-A-N, which stands for [Comprehensive R Archive Network] it will come up as the first thing, and you can go and install R in your computer for no cost, which is very nice for graduate students.

This package, all you need is a vector of impacts, and the variance-covariance matrix associated with those impacts, and it will kick out the p-values that are adjusted for the correlations among these tests.

Now, the tricky part of that is getting the cross-equation

covariances for your impact estimates because normally you're going to be estimating several impact regressions, and what you need are the covariance between the impact in regression one and the impact in regression two, and that's what's not transparent.

Now, I will tell you one way to do this that I know works, and there may very well be other ways to do it. But because I know this works, I'll just tell you what it is.

In the software package STATA, there is a command called suest, and if you estimate several occasions sequentially and give them a name—like equation one and equation two and equation three—you can then take suest, and you say 'suest equation one, equation two, equation three.' If there is clustering, you indicate what the cluster variable is, and this command will give you impact estimates; the covariance for those impact estimates, adjusting for clustering, and giving you those cross-equation covariances.

You can then take those impact estimates and that covariance matrix into R, so this is not seamless. You have to actually get it from one package to another so there's some effort here. And it will give you the adjusted p-values.

Now, one thing that's important to understand because a lot of folks are usually doing their analysis using HLM or mixed effects modeling, and this is—this command in STATA, this Suest, although it controls for clustering, it is coming from a sort of different philosophical background from HLM. It's using something called generalized estimating equations, which is not—it's a different approach, but it does control for clustering. And

I think Peter has a paper that he's working on or is in review that looks across a lot of different past studies and compares what you would get if you use these two different methods, and I believe he found that they are actually very similar in the context of the studies that he looked at.

So you should just be aware that it's a sort of philosophically different approach to analysis, but it's convenient insofar as you can get these cross-equation covariances out of STATA.

The next approach is resampling, and specifically bootstrapping, although it doesn't have to be bootstrapping. There are other concepts of resampling. This is really based on a book by Westfall and Young from 1993 that is a really nice book actually. It's pretty accessible as these things go, and, you know, I recommend you pick it up.

This allows for not only correlations between the outcomes, but it also allows for more general distributions of those outcomes. One type of distribution that they talk about in particular are skewed distributions. They worry that if you have a heavily skewed distribution, that the adjustments that you would make without taking that into account could be a little off target.

One of the really key points that Westfall and Young make is that resampling in the context of bootstrapping needs to be done under the null hypothesis, and when I describe the algorithm, I'll describe what that means, and here's that slide. So it's coming soon.

I'm going to focus specifically on something called a homoskedastic bootstrap algorithm. There are actually several algorithms in the book, one of which is a heteroskedastic bootstrap algorithm. And then

there's another one which is a rerandomization algorithm. But this one is fairly straightforward to implement, and it's generally consistent with what we would expect.

So we'll focus on this. The idea here is that you start out calculating your impacts and t- stats using the original data as you normally would. So you might have multiple regression equations, an impact from each one, you get the t-stat for each of those, and you put it aside to save it for later.

Next, you define Y-star as the residuals from these regressions, and this, defining it as the residuals, keeping these residuals, what we're going to do is we're going to bootstrap from the residuals rather than from the original outcomes, and this is what you have to do in order to follow that guidance of resampling under the null hypothesis.

The null hypothesis, remember, is that there are no impacts. We want to know what the distribution of t-stats is under that null hypothesis, and so by sampling from residuals, we are essentially differencing out any impacts that we observe so that there are now no impacts in our outcome data. Y-star is not affected by the intervention because we've differenced out those effects, and so when we resample from Y-star, we are resampling under the null hypothesis of no effects.

So something that we need to do now for at least 10,000 times—and really 10,000 times is probably not enough because we're using bootstrapping here not to get a standard error, which is, at least in my case, what I was accustomed to using bootstrapping for. We're actually using the

bootstrapping to get a p-value, and you need a lot more applications to do that.

So you probably want, you know, 50,000 times if you have the time. So what you need to do is you need to randomly sample, and I should say that I'm assuming that what we have here is random assignment of schools to treatment and control and students clustered within schools. So what you need to do is you need to randomly sample a set of schools with replacement from Y-star. Then you need to randomly assign those sampled schools to treatment and control groups in the same proportion as the treatment and control groups exist in your original data.

You then need to calculate impacts using this resampled data, and you need to save the biggest absolute t-stat from the groups of t-stats at that replication. So you're going to save the biggest absolute t-stat. So you repeat this 10,000 times, and what do you get? Well, you have 10,000 maximum t-stats, so you have a distribution of the biggest t-stat, and you can see where my original t-stat or set of t-stats fall within that distribution of maximum t-stats.

And if it falls really high in that distribution, if I'm in the right tail of this distribution of maximum t-stats, then I have a statistically significant impact, and if it doesn't, I don't.

All right. I'm going to pick it up. Amy let me know my time here. So this is an example of what a distribution of maximum t-stats might look like. We've got two columns of individual t-stats across—I said 10,000, in here I did [9,000]. So I wasn't really following my advice—but if you

calculate. So I've got two original t-stats, and I've got a distribution of the maximum t-stat.

The adjusted p-value is the proportion of the time that my original t-stat is bigger than the maximum t-stat. I figured that out there. You'll have to take my word for it.

Implementation of resampling. There is a procedure in SAS called MULTTEST that implements resampling, but only for nonclustered data.

A simple approach would be to simply aggregate your data to the school level and apply MULTTEST. A more complex approach would be to actually write a program that implements the algorithm that I described, and then that would give you the exact thing.

So I think what we're probably most interested in here is comparing these methods, and so what I've done is I've generated some data where we have three outcomes, and I've looked at three cases of different levels of correlation, .2, .5 and .8 between the different outcomes.

I had looked at both normally distributed and skewed data, but my results here are really focusing on the skewed data because that's what Westfall and Young said would be where the resampling might show some advantages, and I looked at four types of comparisons. I'm comparing these methods in terms of their ability to control the FWER because, of course, they need to all do that in order to be valid.

I'm comparing them in terms of the magnitude of that factor, which is the adjustment factor to your original alphas. I am then comparing them in terms of what the minimum detectible effect size would be for a study

that would be based on these. And then, so that's sort of like should I use these methods at the design stage of a study is the purpose of the MDEs.

And then I'm looking at a sort of what I'm calling a "goal line" scenario, which is when you're analyzing your data and you have alphas and you're close to significance, can one of these methods sort of push you over that goal line where others might not be able to?

So this slide simply shows that no adjustment does not control the family-wise error rate. I calculated these family-wise error rates using Monte Carlo simulations. I'm not going to describe the simulation. No adjustment clearly does not control the family-wise error rate. Remember, we want these values to be .05, and no adjustment is significantly larger than that.

Bonferroni clearly does control the FWER, but you can see when correlations are high, when you have a correlation of .8, Bonferroni goes a little too far. It's a little too conservative.

And then you can see generalized Tukey and bootstrap. When the correlation is high, they get it just about on the nose: .49 and—or .049 and .051.

So in terms of controlling the FWER, all of these methods are valid. It's just some are more conservative than others.

When you look at the magnitude of the adjustment to your alphas that you have to make, you can see that when the correlation is low, there's not really an advantage of the generalized Tukey or the bootstrap relative to Sidak, but when you have very high correlation of .8, then there's a pretty noticeable difference.

The adjustment to your alphas with Sidak is about 2.85, and the adjustment with generalized Tukey or bootstrap is about two. So that's a pretty noticeable improvement.

What does this mean for MDEs? If you're calculating minimum detectible effect sizes at the design stage of a study, trying to find out how many schools and students you need, do these more complicated methods make a difference? And we can see that going from Bonferroni to Sidak, so where you're going from an approximate control for independent test to an exact control for independent test, you get a little something.

You go from—in the scenario I outlined here—you go from an MDE of .25 to .24, which isn't great, but it's something.

And then if you have really highly correlated outcomes, you get a little extra something going down to .23. So it makes a little bit of a difference in minimum detectible effect size calculations, but it would not be the end of the world by any means if you use Bonferroni to calculate MDEs. You'd be a little conservative, and it's not bad to be conservative when calculating MDEs.

I don't recall being in a Technical Working Group where I say I had more power than I thought, and the TWG members said, oh, really, what went wrong?

[Laughter.]

DR. DEKE: So if you want to be conservative, using Bonferroni is not such a bad thing.

Now, the "goal line" scenario is when you've actually analyzed

your data and you have an impact that is close to statistical significance. In this situation, getting a little extra power from a more precise, more exact, more correct adjustment can make a difference.

So we see here in the situation I cooked up that Sidak gives you a p-value of .054, which is close to statistical significance, but close does not count in this particular game.

When we have higher levels of correlation, the generalized Tukey and the bootstrap can push you over the goal line into statistical significance. So, in those scenarios, it can make a worthwhile difference.

So just to go over a little summary of the whole thing here. Peter laid out the overall multiple comparison guidelines: specifying confirmatory analyses in study protocols; we want to have defensible outcome domains; conduct hypotheses tests on domain composites. And we've seen examples of how the Regional Educational Laboratories have been able to implement these.

But we need to make adjustments for the between-domain scenario, particularly, especially, when an intervention is going to be deemed effective if any one of those domains are affected by the intervention, and we've outlined some methods that can do that, and we can see that at the design stage when calculating MDEs, it's not critical to use the more complex things, but when you're actually analyzing your data, the more complex things can get you over that goal line.

And here are some references and contact information. There's a reference to Peter's paper on the guidelines which is available at IES's

website. There's the reference to Westfall and Young, which is a very helpful book on resampling and gives a wealth of information. I only scratched the surface of that. And, of course, our e-mails.

Thank you.

[Applause.]

DR. FELDMAN FARB: Okay. Many thanks to Peter and John.

We'll go to our first discussant now, David Judkins, senior statistician at Westat.

MR. JUDKINS: Well, thanks for the invitation, Amy. So how many minutes do you think you want to give me now?

DR. FELDMAN FARB: About seven.

MR. JUDKINS: About seven. Okay.

[Laughter.]

MR. JUDKINS: First of all, I'd like to commend the authors on their fine work. I found nothing to disagree with.

I would like to spend my time talking about the nature of this confirmatory versus exploratory analysis. You know, how to know which one you're doing? Or how to group outcomes, which is one of the recommendations? How to drill down—I'll define that later—and the utility of single-dimensional summaries and multi-dimensional outcomes.

I want to thank one of my coauthors at Westat, Andrea Piesse, for helping me out and reviewing these.

Of course, my remarks are personal.

So I want to introduce you all—maybe you've heard of him—to G.E.P. Box. I haven't read any work by him directly on multiple comparisons or false discovery control, but he's written elegantly about the nature of discovery and the use of statistics in that process.

An understanding of his work will help researchers distinguish between exploratory and confirmatory analysis in their own work.

He wrote this great article called "Statistics for Discovery." And it's in the 2001 *Journal of Applied Statistics* so it's based upon his 2000 Deming Lecture, which I was privileged to see, and I'm going to borrow liberally from that here.

His main point is that knowledge development is an iterative process. It alternates between induction and deduction. In the inductive phase, we use new data to improve current models. In the deductive phase, we design and conduct experiments that test the logical consequences of those improved models.

This is not his idea. He acknowledges that Francis Bacon discussed the iterative nature of knowledge development at the beginning of the Age of Enlightenment, and Steve Stigler told him that the idea goes back to 1200s, the founders of Oxford, and they talked about it all the way back to Aristotle.

So, what is this iterative nature? Box has got a charming illustration which I've reproduced word for word. So let's start off with the model: today is like everyday. The deduction: my car will be in my parking space. Data: it isn't. The induction: someone must have taken it.

So I have this new improved model: my car has been stolen. Deduction: my car will not be in the parking lot. But I have data: no, it's over there. So I'm going to induce: someone took it and brought it back.

[Laughter.]

MR. JUDKINS: I have a new improved model: a thief took it and brought it back. Deduction: my car will be broken into when I go over and look at it. Data: no, it's unharmed and locked. So my new induction is that someone who had a key took it.

[Laughter.]

MR. JUDKINS: Aah, maybe my wife used my car. Deduction: she's probably left me a note. Data: yes, here it is.

So I think the entire research community is constantly going through these inductive and deductive phases. Of course, in education research with 5 or 10-year studies, it takes awhile to go through these, but we do.

Box also talks about the role of the judge versus the detective. In the trial, there's a judge and jury before whom, under very strict rules, all the evidence must be brought together at one time, and the jury must decide whether the hypothesis of innocence can be rejected beyond a reasonable doubt. This is very much like a statistical test.

However, the apprehension of the defendant by a detective will have been conducted by a very different process. And the approach of the detective closely parallels that of the scientific investigator.

So how do we fit randomized trials into this paradigm of judge

versus investigator? Well, let me back up a little bit. Randomized trials is, I believe, the name everyone in education research favors for experiments. Much of the tradition for how to run them and analyze them comes from the fields of medical interventions, devices and pharmaceuticals, where, of course, they're known as randomized clinical trials. So we've just dropped the clinical here.

But so we're not taking over the whole tradition because we changed the name. It's not clinical work. So what aspects of that tradition are appropriate in education research?

Well, you know, CRTs are very important in the regulation of pharmaceuticals and medical devices in America. And FDA panels, I'm going to say, those where they invite people from outside the agency, are like Box's juries, and the FDA administrators who make the final decisions are like the judges.

But, of course, there's a huge set of investigators at the drug companies working to synthesize new drugs and other companies to develop new devices. But there's this severe administrative and legal separation between the two operations.

Now, the education researchers I think, you know, we wear both hats, and so it can be hard time figuring out when we're supposed to be judges and when we're supposed to be investigators. There's not separate bodies to do the two operations, but I think this determines to a large extent whether formal control or family-wise error rates is appropriate and, thus, whether adjustments must be made for multiple comparisons.

So I would apply the sort of work that Peter and John have talked about if I'm wearing the judge's hat. So, you know, when should we be wearing a judge's hat and when are we investigators?

I would say that we should treat analysis as a confirmatory analysis if there's a good chance that the findings will become accepted knowledge for years to come. You know, if there's not very much opportunity to contest it.

I also think that there's fairly strong danger of exploratory analyses being mistaken for confirmatory unless we're all very clear in the language that we use when we give results of exploratory analysis.

So, you know, IES has sponsored the What Works Clearinghouse, and the title suggests that all the guidance to be found there is very solid and reliable, and thus I think that requiring FWER control for entry into the What Works Clearinghouse is very appropriate.

But then how do we facilitate the induction phase? How do we work to improve the models that for the most part are still very primitive in education research? A What Might Work Clearinghouse?

[Laughter.]

MR. JUDKINS: I think we need, you know, some place to put the findings that don't pass that same level of control where they can help form people's intuition. We can perhaps also report findings from poorly controlled observational studies. So it would be a resource for experimenters, not for implementers.

And I've got more, but maybe that was 7 minutes. Do you want

me to stop?

DR. FELDMAN FARB: Yes. Thank you. Thank you, David. Those were great points.

[Applause.]

DR. FELDMAN FARB: We have our second discussant, Jeff Smith, professor of economics at University of Michigan.

DR. SMITH: PowerPoint and I don't get [along] so I'm using a PDF file that I made in Word.

[Laughter.]

DR. SMITH: Anyway, good work, thoughtful work. David Judkins and I are completely in agreement on these aspects. Clearly written report. Useful results. Thanks for the invite. It's fun to be here. This is my first time and it's quite interesting.

Let me get right to it since we don't have a lot of time. In some ways, David set me up here by discussing, wheeling in Steve Stigler, whose class I took at Chicago, and the whole nine yards with all the philosophers and history and everything like that.

That's part of the fun of this part of statistics, I guess, is that there are big philosophy aspects here. Classical statisticians think about this stuff very differently than Bayesian statisticians do, and even with the classical field, there are sort of different viewpoints about how to think about this stuff.

At the same time, you know, philosophy is a different thing than science, which, you know, Institute of Education Sciences. It's not the

Institute of Education Philosophy. So I don't know.

Anyway, this is all about avoiding false claims if the treatment has an effect on something, an impact on something. And when is that ever an interesting question? I think it's an interesting question in the sense that it helps to salve the egos perhaps of the researchers or the people who develop the treatment, but I'm not sure it's all that often ever an interesting research question.

And what came to mind in thinking about that is, you know, I read a lot of student papers, both undergraduate and graduate student papers, and oftentimes—and I don't know how this happens, especially when some of these folks have taken undergraduate statistics from me, but they become—

[Laughter.]

DR. SMITH: —deeply enmeshed in telling me about science and statistical significance levels and completely ignore magnitudes of coefficients.

And it seems like there is kind of an analog here that we've become obsessed with: can we get some stars somewhere instead of thinking about what's actually the question that we care about?

And so I want to wield some economics into this discussion, and say, okay, isn't the thing that we actually care about the estimated difference between the benefits and the costs of treatment? Right? That's one number.

No multiple comparisons problem there. That's one number with a standard error that we can sort of sidestep all of this philosophy, all of this worry, all of this fine work that these folks did, by just focusing on this thing

that, you know, from the sort of green eyeshade economist practicality kind of perspective is what we ought to be worried about, which is, does this thing pass a social cost/benefit test?

If it does, let's keep it around. Or let's do it if we're not doing it already. If it doesn't, on to the next thing. On to the What Might Work Clearinghouse. I like that a lot. That's good.

[Laughter.]

DR. SMITH: Why is this nice? You don't have to do all this philosophy about what constitutes a domain, and I've sat in plenty of Technical Working Group meetings where we've sort of fussed about, you know, are these two domains or one domain, and you know is it the first letter that defines a domain or whatever?

You can come up with lots of schemes for defining domains, and that's all very interesting in a sort of coffee house and beret sort of way, but maybe we'd like to be able to—

[Laughter.]

DR. SMITH: —to skip that stuff. We don't have to worry about all the philosophy about multiple comparisons adjustments.

And virtue number three here, my list of virtues, it's the thing we actually care about.

Now, of course, I'm an economist, I have to say on the other hand. So my next slide is what's wrong with my proposed approach? Let me say that, particularly in education research, it's hard to monetize some of the outcomes, right? For a benefit/cost analysis everything has to be translated

into a common unit so that it can all be added up and we get this magical one number. Well, some things are hard to monetize. That doesn't mean we shouldn't try, but it means that we're going to end up arguing a bit about how we monetize some of these things, and oftentimes in education research, we're not able to measure all the outcomes we might like to measure.

We might think that some particular treatment is going to affect not only test scores, but maybe it's going to affect something at home, or it's going to affect behavior outside of school, or whatever things that we can't measure, and so they're omitted from the cost/benefit analysis.

This is not specific to education research. It's true in active labor market programs and other contexts as well. And most IES studies, unfortunately, have relatively short follow-up periods, and so there are always these lingering issues, which I guess are going to be discussed to some extent in the session this afternoon, about, well, maybe if we just waited longer, the things would really work.

And, you know, maybe so. But that's a limitation here, too. To do a full social cost/ benefit analysis, you would like the whole stream of costs and benefits, and then of course there's some philosophy that sneaks back in when you pick the social discount rate for these things, as well, and there's a whole literature about that. It's arisen mostly on the context of the climate change stuff.

You know, are our future generations going to be richer anyway so that we really don't have to worry about them or they are going to, blah, blah, blah, all that stuff, and that's—the philosophers are right back there in

the room again. Hard to get rid of them.

So that's my basic thing. Another way to think about that is I'm suggesting that there really is, this is a different way of aggregating into a single domain where the domain is now, not the variables themselves or groups of variables. It's the benefit/cost difference.

And I am extremely positive and often rhapsodize and I think bore people about how wonderful the work that IES has done is, which is very unusual for me being a Chicago economist. But one limitation it has, I think, has been a bit too little attention to cost/benefit analysis.

So this is my last slide, and then my 7 minutes of fame are done, I guess. Heteroskedastic case is important here—these are smaller comments—especially if you have heterogeneous treatment effects.

It wasn't clear to me why you wanted to use the residual resampling rather than plain old boring resampling. Maybe we can talk about that at lunch or something.

STATA can draw bootstrap samples, too. They have a command called "bs"—

[Laughter.]

DR. SMITH: —for drawing bootstrap samples. And the other thing that reviewing this, the slide show, and thinking about this whole research program led me to sort of a broader question that I think is quite interesting.

I was discussing it actually at dinner last night, which is—this is slightly more dismissive than I intend—but there's a sense here in which the

Regional Educational Labs are sort of doing research by number. There used to be these things called "Paint by Numbers." This is sort of "Research by Numbers." And it sounds derisive, but it's not meant to be.

What has gone on here is that IES has defined a very clear format within which to conduct experimental analyses and within which to report the results of the experimental analyses, and that's a very interesting exercise to think about in and of itself. And we might like to think at some point—we couldn't randomly assign this—but we might like to think, nonetheless, about evaluating that enterprise.

Are we better or worse off having imposed this very firm structure on the RELs relative to what would have happened if we hadn't? Or are there at least certain dimensions of it there where we might? They've given a lot of flexibility in terms of the subject area but very little flexibility in terms of exactly how the research is done.

That's a really interesting question. Maybe we should be doing graduate students like this—I don't know—giving them very firm straightjackets.

Thank you for your attention.

[Applause.]

DR. FELDMAN FARB: Okay. We do have some time for questions. Please step up to the mic and identify yourself.

MS. CONAWAY: Sure. My name is Carrie Conaway, and I'm with the Massachusetts Department of Elementary and Secondary Education, and I have a question for John.

I noticed toward the end of your presentation, you had a little note about the assumptions you made when you did the power calculation stuff, that it was 60 schools and 60 kids in a school. And I'm just thinking about the types of programs we implement at the state.

Probably the largest school redesign program I can think of is our Expanded Learning Time Program, which has 26 schools. And I know that the number of schools really affects the power quite a bit in studies.

I'm curious if you had done the same calculations for a smaller number of schools, how different your results might have been?

DR. DEKE: So there are two aspects in which you might imagine that those parameters of the number of schools or the number of students might matter.

One question would be does this affect the magnitude of the adjustment for multiple comparisons? And the answer to that is no, it doesn't. The size of the study doesn't really affect the magnitude of the adjustment for multiple comparisons.

The thing that affects the magnitude of the adjustment for multiple comparisons is simply the number of comparisons and the degree to which the test statistics are correlated. So sample size doesn't come into that aspect of it.

But in terms of where those MDEs in the table that I showed are sort of centered, you know, the unadjusted MDE, I believe, was .21. If you had a much smaller sample size, then, of course your MDE would be a lot higher. So it would be relevant in that sense, but in terms of the relative

performance of these different methods, it's not going to make a huge difference.

DR. BRAUN: Henry Braun, Boston College.

I wonder if this idea of the What Might Works Clearinghouse is an argument for using the FDR as an exploratory tool? Because I think there's a lot of empirical and theoretical research that shows the FDR has good properties in terms of identifying possible stars, if you will, while controlling error rates under a fair variety of models. Perhaps not under all models of dependence.

And so maybe we should be less rigid in thinking about what are the appropriate controls for simultaneous inference when we're in more of an exploratory mode than when we're in the confirmatory mode.

I also would just point you to an article that Tukey and Jones wrote, I think around 2000, which provides a very strong intuitive argument for using the FDR as kind of an adaptive approach, a sort of intuitively adaptive approach to simultaneous inference control.

And then lastly, I think, also Tukey argued that we should prefer methods that allow us to construct confidence intervals rather than simply p-value things wherever we can for exactly the reason that significance testing by itself provides limited information in terms of real world considerations.

DR. SCHOCHET: Those are excellent comments. I agree with all of them. As far as exploratory analysis and what sort of multiple comparisons adjustments should be done, there are no guidelines on that. That's strictly up to the—I think there's some latitude among the researchers whether they want

to do a multiple comparisons adjustment or not. There are no guidelines on that.

But I agree that for an exploratory analysis, the Benjamini-Hochberg type approach does seems to make more sense, but for a confirmatory, you don't want the sort of how much you're reducing the Type I error to depend on how many significant impacts you find. It doesn't seem to be a confirmatory test. That's why we're focusing on the family-wise error rate. But your points are all well-taken.

DR. REARDON: I'm Sean Reardon from Stanford.

I just have a comment on Jeff's comment, and while I like the elegance of your sort of solution to this, I think we lose something really important by focusing on the cost/benefit confidence interval rather than the thing, and that is, this is after all the Institute of Education Sciences, not the Institute of Educational Pragmatism—

[Laughter.]

DR. REARDON: —and it seems to me that the cost/benefit thing is an important thing to know from a very pragmatic point of view, but it doesn't produce scientific knowledge. If all we report at the end of the day was cost/benefit analysis, we wouldn't know sort of why things work or how they work as much as we could if we also look at these other things.

I think we ought to look more at cost/benefit stuff, but to give up the actual measured outcomes and reporting confidence intervals on them with some kind of adjustment I think loses a big part of what we actually care about doing here. So that's all I'd like to say about that.

DR. RANDEL: I have a question about the distinction between—oh, this is Bruce Randel from McREL.

The distinction between the "within a domain" and "between a domain,"—there may be times when conceptually you have something that you think is a domain. You have multiple outcomes in that domain, but it might not factor out that way, and so would you then, what would you do with that? I mean would you call those separate domains and then do the resampling?

And sort of a related question is, you talked about the resampling with high correlations, like .8, but at what point are you getting to the point where you would say, well, the correlations are .8, isn't that one domain? And so how do you balance kind of between those two?

MR. JUDKINS: Well, since my discussion was cut short, and one of the things I had—

[Laughter.]

MR. JUDKINS: —was about forming groups, I have noticed that there's tremendous resistance to collapsing into small domains. Everyone seems to feel like in education research a lot of assessments are published by particular researchers or companies, and there seems to be a great deal of reluctance that I've seen to average different ones together, but I think that it's something that we need to struggle with and look at the context. I agree about the .8. Those to me are in the same domain by that time.

Yeah, like, is receptive English vocabulary and expressive English vocabulary really separate domains? Well, I guess it's in the context. If you're talking about a broader school that also teaches say fencing and

motorcycle mechanics, then I think we all agree they're in the same domain. So context matters.

DR. SCHOCHET: Yeah. I might add, too, I think that point just furthers the use of the resampling because if you use the resampling, and they are .8 correlated, it doesn't in some sense statistically matter so much if they're the same domain or separate domains, whereas, if you use the Bonferroni, which assumes independence, then it matters a heck of a lot if you put them in separate domains. You lose a lot of power.

DR. FELDMAN FARB: Okay. We have 1 minute left so you'll be our last question.

DR. ABER: Larry Aber from NYU.

Linking the Secretary's comments this morning with Peter's call that theory needs to be used to set up those confirmatory hypotheses, any parting thoughts about the role of theory in integrating things across studies and in relationship to the methodological work you're doing right now?

Because it seems to me that theory is underdeveloped and needs to be developed as powerfully as the methods.

[Pause.]

DR. ABER: No theorists?

[Laughter.]

MR. JUDKINS: I agree that theory is underdeveloped and we need to focus more on the inductive phase of it. I think a lot of IES research has been focused more on the deductive side. At least, that's the part I've been involved in, and how do we facilitate the new theories, because a lot of

what we've tested is shown to be ineffective.

I did one study where the theory strongly guided the measurement of adherence to guidelines, so it was like, how closely did the teachers do what they were supposed to do? And we found no relationship between fidelity and the outcomes. So the theory was totally wrong, I would say, and that's just one example.

DR. SCHOCHET: I mean I might just add that, you know, all of the 25 REL studies, they've all had some sort of a conceptual model underlying, and they need to justify that model in order to get IES funding. But whether it pans out or not is, you know, a function of underdeveloped theory, and Larry, you're much more able to speak to that than I think any of us.

DR. ABER: If that's the case, we're in trouble.

[Laughter.]

DR. SCHOCHET: Okay.

DR. FELDMAN FARB: Okay. It's time to draw this to a close so we can all get back. We thank you for your time this morning. Enjoy the rest of the conference.

[Applause.]

[Whereupon, at 11:35 a.m., the panel session concluded.]