INSTITUTE OF EDUCATION SCIENCES

FIFTH ANNUAL IES RESEARCH CONFERENCE

CONNECTING RESEARCH, POLICY AND PRACTICE

CONCURRENT PANEL SESSION II

"Design and Analysis of Single-Case Research"

Tuesday, June 29, 2010

3:05 p.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. BUCKLEY: Good afternoon and welcome to the session on the Design and Analysis of Single-Case Research.

I want to say just a couple of administrative notes first. As you can tell by the bright lights, this session is being videotaped. So, please, as you would for any session you attend, make sure you turn off cell phones, pagers, those kinds of things, so they won't interfere. And by entering this session, you are consenting to be audiotaped and videotaped. So, if you're not comfortable with that, you might need to choose another session. I think those are all my administrative notes.

I am pleased to be able to present this panel on single-case research. Single-case research has a strong history in special education, in psychology and behavioral research. And IES really views single-case design as one of the important methodologies in the toolbox, if you will, for researchers, particularly in special education, to

help us understand what works for whom and under what conditions.

And despite the strong history, though, there are continuing questions about what makes a really rigorous single-case study and what are the most sophisticated analyses we can use to understand the results of single-case research within a single-case study as well as across single-case studies, and across single-case research and results from group designs.

Again, I am very pleased to have this panel here today, which represents a lot of the efforts that IES is making to try and answer those questions and to be able to advance our knowledge and understanding of what makes a really rigorous single-case study and what are those best analyses that we need to be conducting around single-case research.

Tom Kratochwill is going to be our first speaker, who is actually presenting two presentations. He is going to present information from the What Works Clearinghouse. He was the chair

of the group tasked to come up with standards that the What Works Clearinghouse will use to evaluate the quality of the evidence presented in single-case research, and he's also going to present some of the results from his IES grant that he has with Joel Levin at University of Arizona, looking at some of these issues of how to improve the rigor of the design of single-case research.

And then we have Larry Hedges, who is going to describe his work with Will Shadish in a recently funded IES grant on a d-estimator for single-case research. I know many of you in this room are on the edge of your seats wondering if we have answers for effect-size in single-case research, so Larry will talk about some of their work in that area.

And then we'll have Rob Horner, whom many of you, of course, know—has a long history and is a leader in the single-case design and research world.

Let me introduce Tom Kratochwill, who is the Sears Roebuck Foundation Bascom Professor at

the University of Wisconsin-Madison and the Director of the School Psychology Program.

He's also the Director of the Educational and Psychological Training Center, an interdisciplinary unit for clinical and applied training. He's also the Codirector of the Child and Adolescent Mental Health and Education Resource Center.

He has numerous career awards for research contributions and for teaching. He's a past president of the Society for the Study of School Psychology and Cochair of the Task Force on Evidence-Based Interventions in School Psychology.

He's also a member of the APA Task Force on Evidence-Based Practice for Children and Adolescents, and many of you, I'm sure, have read his books and articles on single-case research.

Welcome, Tom.

DR. KRATOCHWILL: Thank you, Jackie.

What I'm going to do is, first, as Jackie mentioned, present some of the work that we did as a task force for development of standards for

single-case design for the What Works Clearinghouse. That work was managed by Mathematica, and we had a panel of individuals who we worked with over a period of a couple of years.

In fact, this is truly a work in progress because, as we speak, some of the standards are still being revised and developed. I have been told that by the end of the week, this week, we will probably be able to post some of the standards documents and the corresponding white paper on the WWC Web site. Stay tuned for further developments in this area.

Some of the issues that I'll present today, actually, there have been some revisions in, and I'll point this out because when the PowerPoint slides were sent in, we actually had not resolved some of the issues, and I think once these standards are posted on the Web site, we'll probably be revising even some more because there probably are some controversial issues, which brings me to a point I want to emphasize: anything controversial that comes up today, please direct

your comments to the discussant.

[Laughter.]

DR. KRATOCHWILL: Rob has indicated that he's very well equipped to handle these kinds of things and has spent his career actually fielding difficult questions.

Anyway, this is our panel, and it was a great group to work with. We developed consensus on most things that we addressed in the panel, but there were a number of issues that I'll point to later this afternoon that we were not able to address and did not address, and perhaps these remain for further development.

What we were asked to do is to work on standards for single-case research design, and one of the points I want to emphasize is that when we talk about single-case design, the context is that we're looking at the tradition of single-case intervention research. We were not looking at qualitative case studies, for example. We were looking at designs in which an intervention is being tested—interventions for social, emotional,

behavioral, academic issues—that would be of interest to educational researchers and practitioners.

And so this slide kind of features some of the defining features of a single-case design that we looked at. This seems like a pretty obvious point, but in the world of research beyond what we think of as single-case design, there's actually quite a few different terms thrown around and conceptualizations of what this research involves, especially in the medical area. This is what we focused on.

Traditionally, there have been three general classes of design, and this is what the standards panel focused on: the ABAB design, the alternating intervention design, and the multiple baseline design.

This domain way of looking at these designs is the process that we used to organize our standards and think about how we conceptualize standards for research in the field of education and special education.

As we developed standards, we were asked to address the more traditional threats to internal validity, and, in a white paper that will be forthcoming, we also took on the task of indicating how single-case design researchers can address internal validity threats in their research, and, there's quite a comprehensive review in this area.

It's an area, for example, that Will Shadish had a lot of input into, and it was an interesting discussion primarily because we're not talking here about randomized clinical trial research. We're talking about research in which replication is a primary feature to address internal validity standards.

The types of questions that single-case designs might answer:

Overarching, was the intervention effective—which intervention is effective for a particular case?

And depending on what the researcher's question is, these are some of the classes of traditional questions that might be addressed in

the type of research standards that we developed.

One of the more interesting aspects of the panel's work was thinking about how we craft standards for looking at designs versus those for evidence, and, in many ways, this was a feature of the panel's work that I think will have an important impact because what we tried to do is separate design standards from the standards in analysis of the data.

For example, you might have a great design, but the results would not demonstrate any particular effect, and what we're truly trying to look at here is standards that would help researchers address some of the issues in publishing negative results or bias in publishing.

In other words, if you have a great design, but you didn't find that something works, that's important evidence for the field, and it kind of fills in with the momentum that the evidence-based practice movement has had.

In doing our work, we paid attention to both design standards and visual analysis standards

for outcome variables that were associated with the experiment.

The criteria that we looked for single-case designs that meet evidence standards is listed in this particular slide.

I'm going to go through each of these very briefly and just indicate what the panel decided, and, again, in some of these, there are still some issues that remain to be resolved.

First of all, the independent variable must be systematically manipulated, and we were interested here in looking at research in which the investigator manipulates an independent variable and looks for the effects on an outcome variable.

More traditional passive observational studies which do not involve active manipulation of an independent variable would not typically meet the standards.

We were very interested in looking at the current literature in terms of how outcome variables are measured in single-case research, and some of these criteria reflect the way that single-

case design is done:

Measurement occurs over time; a typically inter-observer agreement is reported. We indicated that it must be assessed on each outcome variable in every phase, and there should be measurement for at least 20 percent of the sessions distributed across all conditions of the study.

This has turned out to be a fairly rigorous standard, and remember that these standards are being applied to research that has already been conducted. One might think of this standard as helpful as a guideline for future research, but, in fact, a lot of the research is really already in existence, and, hence, we're going back and applying these criteria.

I've been working with the people who are beginning to do pilot reviews, and they are finding difficulty in people meeting this standard of inter-observer agreement. In some studies, the statistic is not reported. It's unclear what phases inter-observer agreement was conducted in, et cetera, et cetera.

We also wanted a standard in which we looked at attempts to demonstrate an intervention effect. We picked the criterion of at least three attempts to demonstrate an intervention, and in our traditional, or in the three domains that I mentioned earlier—the ABAB design, multiple baseline, and alternating intervention design—would meet those criteria because there are at least three attempts to demonstrate an intervention effect. At least, that's the standard that we were invoking.

There are designs that do not meet that standard. Examples are AB, ABA, BAB designs, and other designs in which there is not this attempt to replicate the intervention effect.

This is an example actually from some work that Rob Horner and his associate have published, demonstrating the example of where the intervention effect is replicated, and you can see in the bottom of this graph that there is a first demonstration, second, and third demonstration of the effect. That particular design would meet the criteria because

it has this basic replication standard.

Similarly, a multiple baseline design with three replications—this one has four—would meet that standard as well. Two replications would not.

The next criterion is really a subpoint, where we looked at some exceptions with certain designs, and there may be studies in which there are fewer than three or four data points in a phase, and we decided that there may be standards met but with reservations.

And there are exceptions in certain designs that are used in applied—educational research, such as in the alternating treatment design, randomized designs, and brief functional assessment.

In fact, in some randomized designs, there may only be one or two data points per phase. In some designs with brief functional assessment, one data point. We built in some safeguards to not throw those kinds of studies out because they were designed to look at a different component of replication.

What we did do is build in some criteria for the number of replication attempts and decided on five.

When we looked at the standards for analyzing the data in single-case research, there was pretty strong consensus that we would need to meet evidence standards through visual analysis. Despite rhetoric to the contrary, there have actually been very few statistical applications in single-case design. I studied that back in the late 1970's, and, actually, after almost 30 years, the picture has not changed tremendously.

There was general agreement that visual analysis would be a good standard to use in this work.

We also thought, though, if we're going to have reviewers look at single-case designs to draw inferences from the outcome of the study, that they would need to be trained in visual analysis so those criteria were necessary, and we borrowed very heavily from some of the work that Rob Horner and his associates have done in terms of training in

visual analysis and the criteria that need to be taken into account in visual analysis of data.

This particular slide shows six variables used in visual analysis and a four-step framework that guides the visual analysis process. Actually, it could be used in the analysis using statistical inference as well. The visual analysis then is applied to all the designs.

One of the important elements of the visual analysis is to think about three demonstrations of this effect at different points in time. We invoked this basic effect, which is change in the dependent variable when the independent variable is actively manipulated, but you need to take into account a variety of criteria that researchers have invoked for doing a good visual analysis, and these include the elements that are listed on this slide: level, trend, variability, et cetera.

When you do visual analysis, however, we invoke the standard that experimental control involves all phases of the study, not just any two

adjacent phases. In other words, the researcher has to look at the entire replication series in the design.

The basic effect is done with adjacent phases. Experimental control, however, involves all elements of the design, and that was an important element.

This breaks apart the four steps and six variables in terms of visual analysis, and I'm hoping that Rob may say a little bit more about this today, especially if there is controversy or people become angry and hostile about it.

[Laughter.]

DR. KRATOCHWILL: We invoke one additional standard for multiple baseline designs, which those of you who are involved in conducting single-case research know, and that is you want stability in the nonintervened series when the effect is demonstrated in one series and subsequently across series of the design, whether units, behaviors, or settings. That standard needed to be revised slightly.

In alternating treatment or multielement designs, we considered magnitude of separation between the two conditions, the consistency of the separation, and the number of data points as important criteria as well.

What does this mean? This is a very, very brief presentation of some of the standards. You're going to be looking at them very shortly and draw some of your own conclusions about how well they serve our needs in reviewing literature in the field of education.

But there is something to be said here in terms of the reliance on visual analysis or visual inspection of data. This is actually a bit of a segue into my subsequent presentation here this afternoon, but I'd like to mention a few things about why we selected visual analysis because there was actually some controversy surrounding using visual analysis as the primary criterion for determining an intervention effect.

First of all, a lot of the research in education, and especially in special education,

school psychology, clinical child, et cetera, where single-case designs are used, has been in the tradition of applied behavior analysis. There has been a very heavy emphasis on visual analysis.

In fact, some of our best work in terms of the criteria for conducting a good visual analysis as well as some of the research surrounding the issues in using visual analysis have come from that tradition.

Secondly, there's also a lack of consensus surrounding statistical analysis of single-case design. Just when you think you've found the ideal solution, someone will come along and find a major limitation either based on Type I error, autocorrelation in the data, possibly flaws in the computer programs that have been implemented, et cetera, et cetera. And so, there isn't wide-scale consensus on what is the best method to use in analyzing the data.

And then, of course, many practitioners who are involved in conducting single-case designs in their practice settings rely almost completely

on visual analysis of the data.

But we were under the assumption that we could improve the judgments of some of the existing literature out there by invoking several procedures that we think would be helpful, and we relied very heavily on the first one, which is structured training in visual analysis, and that typically involves comparing visual analysis outcomes of novices—that is, people who are beginning to look at graphs—and judge the intervention effect, to experts, and we have just really begun that process as part of the training.

Rob Horner and I went to Mathematica this past winter and did a training to the first round of reviewers that are now wading through the literature on single-case design, and, actually, the training went really quite well.

I have to say, however, that there is a paucity of research on the best methods of training and the structure the training should take in terms of guiding our efforts in this area.

We've begun to do—actually, had a

literature review on all the work on visual analysis, and there isn't a tremendous amount to show what is the best method of training and how do you do it. However, we are forging forward with that focus.

There are some other elements that have been used in visual analysis, and that includes some of the protocols, for example, that Tawney and Gast introduced years ago, where you do a fine-grain analysis of the basic effect from phase to phase and then a total analysis of the effect across phases. So, that may be helpful.

There are some criteria out there for training. Wayne Fisher and his associates introduced the dual-criterion method. They applied it to AB-type designs. We recently extended that—Swaboda, myself, and Joel Levin—to a program in which we use all phases of the design. We've extended it to ABAB designs as well as multiple baseline designs.

In the future, that may become an option for training individuals in visual analysis of

data.

There are some rather novel and new procedures that need investigation that I think we did not really address but have promise. One option is to consider—and I think this is one of the more controversial ones—to use randomization in the design structure rather than doing what most researchers have done in single-case research, and that is response-guided selection of when the intervention is introduced. One might define random points to introduce the intervention.

This is actually a point that John Ferron and his associates have recommended, and Todman and Dugard in their work on randomization tests have suggested that procedure. That may be a possible way to proceed in the future, but, again, lacking is empirical evidence on supporting that over the tradition of response-guided methods for determining the point of intervention. Ferron and Jones also introduced this concept of a blind visual analysis procedure from a data analyst who is blind to the hypotheses of the study, which is

an interesting idea and similar to the idea of having two observers who are independent looking at the data and determining the inter-observer agreement score.

And then, of course, there are a number of proposed and current visual and statistical methods of analysis that might be used, some of which have been published in *American Psychologist*, in some of our special education journals, and other places.

These remain to be investigated and certainly compared to visual analysis.

Now, there are a number of issues the panel did not address. How am I doing on time? Okay. I need to speed up in the second presentation. Okay.

Some things we did not address in the panel. We took the traditional single-case designs. We did not look at the concept of randomization as it's applied to single-case design structure. That's something that I personally feel may improve the interval validity of single-case design, and I'll argue that in just a few moments.

We did not look at or recommend statistical analysis procedures for single-case design. We didn't say, for example, that randomization tests are great or the time series analysis was appropriate in the study, et cetera. There's a literature out there, although small, in which various researchers did use statistical tests. We didn't really say much about that.

We ran into some interesting work, which I think Larry will talk a bit more about this afternoon, and that is how do we come up with a single-case design effect-size determination? After all, the purpose of these reviews is to say something about the overall effect of a particular intervention across various programs or settings or whatever variable is of interest, and we did not reach closure on that.

And then, finally, we did not recommend ways of combining single-case design studies with group design research. That remains to be done and probably will be a very complex statistical process, at least in talking with colleagues on the

panel who have done that. It probably will be a significant challenge to come up with that, but one that's very important because the What Works Clearinghouse is obviously reviewing group research, and single-case design will have its own database, which could obviously be used to embellish or perhaps even contrast with the findings of some group investigations. So, I thank Rob for his work on some of the training material.

At this point, I'm going to go on to the second presentation. Jackie is going to jump up and help me do that. There we go.

If you thought that was interesting, just wait for a few minutes. It's going to get even better. Okay?

As Jackie mentioned, Joel Levin and I have a grant that we've had for a few years with IES, and the purpose of the project was to review various methods of analyzing the data. One was to look at design structure in the context of what kinds of randomization schemes might be used in single-case research. Another was a comprehensive

review, a visual analysis of single-case research designs.

What I'm going to be presenting here this afternoon is some of our work on the use of randomization in single-case design. I'm going to go through this, I guess, rather quickly, but to say that if you were interested in reading more about this and getting a good, you know, laugh or perhaps talking with colleagues about how ridiculous this might be, it is published now in *Psychological Methods*, and that is out, I believe, the June issue. So, you can access that. *Psych[ological] Methods* is an APA journal.

One issue is why would we want to use anything random within these designs? We would argue that it does improve the internal validity of this research. In some ways, it elevates it to the status of a randomized control trial, which has been the gold standard not only in medicine, but as it has migrated into education and psychology, we've seen that as a legitimate method to suggest that there is evidence.

And in the case of randomization in single-case design, there is a class of data analytic procedures that might be used that have pretty good statistical conclusion validity to conduct a data analysis. In the absence of that, the tests are, of course, not appropriate.

In our work, we were kind of inspired by some of the writings of Reichardt in a *Psych[ological] Methods* paper that he published in 2006 and work we were doing in reviewing statistical applications in single-case design.

The contribution from Reichardt was interesting because in contrast to the traditional method of only randomizing participants or subjects in a design, he talked about the importance of keeping constant time settings and outcome variables or randomizing these, and in the case of time series or single-case designs, which could be considered a class of time series designs, one might use randomization in the design to have a completely randomized design, and so, that was one of the bases for it.

Now, this is not a new concept. When I first edited a book in 1978, which seems like just yesterday, we were already talking about nonparametric randomization tests, and only a few close friends and relatives picked up the idea and ran with it, and here we are a long time after that, and we're still chatting about it.

The problem in traditional designs, as you know, is that like an AB design, there is no replication, and so, one is left to think about, well, how can I improve the validity of this kind of design? Aside from internal validity issues, what might I do to improve it?

And we can think of random orders and how those phases are introduced. Now, that isn't a particularly compelling argument with an AB design because you have time-related and carryover effects. You have the absence of replication, et cetera, but the idea that you can randomize which condition comes first is interesting, especially if you have two treatments that you're comparing, or you think of the baseline as the treatment that has

been in place for many years.

Now, in the traditional ABAB design, we also have a fixed sequence. So, in thinking about this work, Edgington, years ago, back in the late '60s and early '70s, began to talk about how you might take a design like this and begin to use randomization.

This design, while it addresses some of the validity issues that we're concerned about, you could think of random assignment as applied to the various phases. This is a fairly complex slide, but it shows if you randomize all the orders, it doesn't really get you a whole lot except perhaps in the condition where you've got an ABAB design because you address initial intervention effect, return to baseline, and a replicated intervention effect.

None of the other random orders gets you that replicated effect. But one way to think about this is not to think just about ordering the phases but perhaps thinking about ways of ordering them across a larger number of sequences, such as in

this example in this slide.

You can actually do that with a simple or a block fashion. The block fashion gives you the option of having both phases represented in the block, whereas, a simple randomization scheme may not. For example, you may have three or four or five successive B phases following the initial baseline.

Another way to think about this, if you want to make more complex experiments, you could have an ABAB design with multiple units. Units here refer to cases or classrooms or schools or whatever unit of analysis you're looking at.

You can do that within a simple or within unit blocked variation. Certain designs allow randomization, but, in fact, you may have to make some adjustments in the way these designs are done. And one adjustment is to have an initial warm-up or baseline phase that always precedes all the conditions of the randomized scheme. That is one option for researchers to consider that has been discussed.

You can look at randomization within the context of some other designs that we're likely to consider using and will find in the literature. This Table 8 gives an example of the alternating intervention design under either simple or blocked randomization, and it's a design that's very easy to randomize.

In fact, among the various designs out there, randomization is used more often in this design than most of the other ones.

The alternating intervention design can also be used across multiple units, which gives you an example here of not only the order but possibly randomizing the units as well.

This is another example where you have three treatments in which randomization is used. Again, the variation could be simple or blocked.

Multiple-baseline designs—one of the more common designs that we see used in the literature and, actually, a fairly strong internal validity design from my standpoint—is one that is very capable of being used in a randomization scheme.

Joel and I argued that in 1978 and demonstrated an example where you would randomize the participants to the order of the intervention, and so, for example, this random order happened to come out three, five, two, four, one, and you can see that the intervention is still staggered, but the order in which people received the intervention is randomized. You can also do this across behaviors and settings as well as participants.

What if you only have an AB design? There, we go back to some of Eugene Edgington's work, where you think of not only ordering phases, which I said earlier is problematic, but possibly ordering the point at which the intervention is introduced in the data series.

Rather than ordering the phases per se, you're going to order the intervention point, and you can do that based on 20 observations, as we have here, or 20 phases, or you could do it based on what we call regulated randomization, where you preselect particular intervals to introduce the intervention.

There are compromises statistically but quite a bit of flexibility in doing that. Again, this design can be implemented across different units or participants in a randomized sense. You can not only randomize the point of intervention, but you could randomly assign people to the orders.

Multiple-baseline designs could be done the same way, and you can stagger the interventions, have different points of intervention within a preconceived regulated randomization structure, or not.

How am I doing on time, Jackie? A couple more minutes. Great.

When we think of these designs, we also can look at not necessarily the AB design where A refers to a baseline or absence of treatment but a BC type design, or what we'll call here Intervention X, Intervention Y, and compare two interventions. The interventions can be compared where there are independent points of randomized start times for the intervention or possibly simultaneous start times.

In fact, these different models have been subject to analysis with different types of randomization schemes or tests. You can take comparative AB designs and increase the number of units, expanding on this concept of independent versus simultaneous intervention point.

What do these give us? Well, one of the interesting things about this, it's not just a randomized design issue—although, I would argue that it improves the internal validity by using randomization—but there's a class of single-case designs that can be used in each of these applications, and the statistical test is a randomization test, and there are many now out there capable of being implemented either by hand or by computer programs that can add to the statistical-conclusion validity of the study.

These are, in my estimation, seen as complementary to some of the effects we may want to see and apply it in clinical research. For example, where we invoke social validation criteria, looking at the magnitude of effect or effects of social

significance or meaningfulness.

Nevertheless, small effects might be detected with some of these particular tests in the context of a randomized design.

In investigating these particular designs, we are involved in developing a variety of statistical tests for them, and, to that effort, I thank Venessa Lall and John Ferron who have helped with some of our simulation studies and data analytic techniques.

Further information on this is available. I understand that these particular slides are on the IES Web site, so you can access this, and again, the *Psych[ological] Methods* paper is now published and available.

[Applause.]

DR. BUCKLEY: All right. Thank you, Tom, for those two presentations.

I'd like to now introduce Larry Hedges, who is a national leader in the field of education statistics and evaluation. He's the Board of Trustees Professor of Statistics and Social Policy

at Northwestern University, where he also holds appointments in statistics, psychology, and educational policy.

He is widely published across disciplines. He's an elected member of the National Academy of Education and is a fellow of the American Academy of Arts and Sciences, the American Statistical Association, the American Psychological Association, and the American Educational Research Association.

And Larry is going to talk about his work with Will Shadish and a d-estimator for single-case designs.

DR. HEDGES: Thank you, Jackie.

The very first thing that I want to say is that it's important that the work I'm going to talk about is part of a joint project with Will Shadish and David Rindskopf. But they didn't get a preview of this presentation, and they aren't here to defend themselves, so, if you don't like it, don't blame them.

I'd also like to start by saying I think

that visual analysis is a really important tool in this kind of work and with these kinds of designs. Even though I'm a statistician by training, I think there's a lot that can be learned by looking at the data and by looking at the data in a structured way that is subject to some rules of judgment.

I also think that we know less than we should about applications of rules of judgment in connection with data, and, I think, not only in this area but in many other areas like standard setting, we've been surprised by the properties of judgments, and I think it's important—it's important to study that as a major research methodology problem.

The talk that I want to give today is to describe a perspective that might be used to generate effect-size measures for single-subject designs. I say "a" perspective, and "a" d-estimator because I think there are other ways to approach this problem, and, in fact, there are other people working on this and making good progress on this. I sort of—I should mention that Rob Horner and

Hariharon Swaminathan have done quite a bit of work on a slightly different strategy to solve the same problem.

You should be aware that there are lots of people working on this and making what I think is pretty good progress.

But I'm going to talk about the work that Will and David and I have been doing and what I hope to do in my remarks today is to sketch a program of work—sort of describe the way our program of research works—and I should say we just got started on this. I think the grant was only funded a few weeks ago, so we've made remarkable progress in the last few weeks.

[Laughter.]

DR. HEDGES: And I'm here to tell you about it. I'm going to show some simple results, and so, the reason they're simple is because we haven't been at this long enough, but I think the simple results I have to show you will suggest where you might go and how we might go about it, and we'll talk a little bit—I'll talk a little bit about

where we intend to go from here.

First point is that effect-size measures are useful for representation of results of single studies. Even though I'm a fan of visual analysis, I think that sometimes having quantitative effect-size measures is useful.

Now, having said that, it doesn't necessarily follow that the d-index is the effect-size measure to use because I have had the experience in various areas of research synthesis and various areas of primary research—that it's useful to have effect-sizes that make sense given the kind of data that's collected and the kind of research designs that are used.

I think effect-size measures have an important function, however, in accumulation of results across studies, and this is important because cumulation is becoming conventional in a lot of areas, and I'll mention in education the What Works Clearinghouse, another endeavor of IES, and, also, meta-analysis more generally is a standard for cumulation of research results.

Having said that, it doesn't necessarily follow that you need special effect-size measures—that you need a d-index for the effect-size measure to be useful for those purposes. Accumulation of results, though, is increasingly becoming essential to make research count in some scientific and policy context, and, as Tom already mentioned, effect-size measures for single-subject designs need to be understandable to those outside the community, and it would be highly desirable to develop effect-size measures that were similar to those or comparable to those that are used in between-subjects designs.

If we could accomplish that, that would help to move, put essentially single-case designs in the same ballpark as between-subjects designs, and it would allow the combination of that information in better ways or the comparison of that information in better ways.

Having said that, an obvious question is, well, what do you mean by comparable? You know, what am I talking about here? Can you make that

precise? And, actually, what I'm going to try to do is make that precise in one way. There may be other ways to do this.

But one way you could imagine making the notion of effect-sizes for single-subject designs comparable to those for between-subject designs is to do a thought experiment and to imagine a large set of data that includes enough data elements that you could look at parts of that data set as being a single-subject design or look at only some other parts of it as a between-subject design.

Essentially, we know how to define the d-index in between-subject designs. The idea would be to create an effect-size measure that you could compute from single-subject designs that would estimate the same parameter as the d-index does for between-subject designs.

And if that was the case, then at least in this thought experiment, in principle, we could estimate the same effect-size from either design, and they'd mean the same thing, and so, that's really what I mean by comparable.

And now, I'm going to try to make that a little more precise. And I'm going to ask you to imagine a large data set and call the data elements Y-ij's. So, imagine rows and columns. Imagine a row for each individual and imagine a column for each measurement over time.

Now, I should say at the very beginning, I'll apologize for this study. I'm going to talk almost exclusively today about one of the designs that Tom eliminated as being not meeting evidence standards. I'm doing that not because I'm a fan of weak designs but because they illustrate the essential features of what we would do with more complex designs, and I think this is complicated enough with two phases. It will be easier if we stick to this and, then, trust me, that you can generalize this.

Imagine a situation in which we'll index the rows and columns of my imaginary data table by i and j—i for the individual and j for the measurement—and we'll imagine the first n of the measurements are in the baseline phase, and the

next n are in the treatment phase, and the two n's don't have to be the same. They don't have to be the same across individuals, but it sure makes it easier to write down the notation if you have the same number of observations in the baseline and the treatment phase for everybody.

And now, here's where the thought experiment comes in. Imagine we have this whole data matrix of individuals who have a baseline phase and multiple measurements there and a treatment phase and multiple measurements in the treatment phase.

You could obviously take several of those individuals and treat this as a single-case design. Each time series is its own study in a way, and if you have several of these, you have time series with replication.

Now imagine a slightly different thought experiment. We have all the data. Suppose I randomly choose certain of those individuals, and I declare them to be the control group, and for each of the people in the control group, I choose an

observation from their baseline phase, and for each of the individuals in the treatment group, I choose an observation from their treatment group phase.

That's kind of the idea of imbedding the possibility of both kinds of studies within the same data set. What I would argue is that if we have a structure like the model that I've suggested here, I'm suggesting that within an individual and within a phase, we can think of each individual as having a mean for a phase, and then there's a residual for each different observation within the phase.

And suppose the variance of these residuals is Sigma-squared, so the variance across observations within an individual is Sigma-squared. Each individual has their own mean, and there's variation across individuals in those person-specific means—the Mu-i's—and imagine that the variation across individual means is Tau-squared. Then you begin to get—I'm beginning to put some structure on this now from which I'm going to derive an effect-size.

Now, the way I've set up my equations over here—I guess we don't have a pointer, do we? Oh, is this the pointer? Oh, cool. Okay. How do we turn it on? There it is. Okay. The basic idea here is an observation. You know, an observation in the baseline looks like an individual's mean for the baseline plus a disturbance term.

The disturbances have some variation, which I'm labeling Sigma-squared. Same individual in the treatment phase has an average that is whatever their average was in the baseline phase plus their treatment effect. It's an individual-specific treatment effect. There's an i on it. That's what that means. You know, plus there's a disturbance for each of these things, and I've just made these two variances within baseline and treatment phase the same.

It's not necessary. It just saves having more different symbols up there. But this basic model is kind of a stationary time series structure within either phase of the design. And then, I've added down here the between-subjects structure by

saying that the individual means in baseline vary across individuals, and the variance is Tau-squared. The individual treatment effects may vary across individuals, and they have a mean that's delta-dot and a variance that I'm calling Theta-squared here.

I'm going to try to convince you that d here is the effect-size parameter that flows naturally from the between-subjects experiment that I have derived from my big data structure. Everybody kind of with me? I just want you to note that delta-dot here is the average shift between baseline and treatment. Tau-squared is the variation across individuals on the average, and Sigma-squared is the variation within individuals within a phase.

What that model implies is that the variance across subjects of any single observation within phase is Tau-squared plus Sigma-squared. In other words, the reason why single observations for individuals vary is partly because the average for individuals is different. That's what the Tau-

squared is about, and because each individual observation within individuals varies, and that's where the Sigma-squared comes from.

The mean difference in our hypothetical experiment between baseline and treatment is delta-dot. The variance of the observations within treatment groups, in quotes, is Tau-squared plus Sigma-squared. The d-index we would compute from my hypothetical between-subject experiment is delta-dot over the square root of Tau-squared plus Sigma-squared.

That is the d-index that you'd get in the between-subject design. I argue if we want to try to get a comparable effect-size from a single-subject design, what we got to do is try to estimate d from the kind of data that we would have in a single-subjects design.

Now, that's made more complicated because the residuals aren't independent the way they are in some nice models that we are used to dealing with. Instead, I've assumed that the observations have a first-order autocorrelation structure, so

the auto covariance matrix—that is, the covariance matrix of the residuals within a phase—looks like that.

And that introduces minor technical complications. Remember, to estimate d, all we have to estimate is delta-dot, Tau-squared, and Sigma-squared. Well, the mean difference between phases estimates delta-dot. It estimates the average shift.

To estimate Sigma-squared, that's the variance of any individual residual, turns out to be slightly complicated by the auto covariance structure. You can't just take the standard deviation across observations on the same person and expect that to estimate Sigma-squared. It doesn't.

The estimation of Tau-squared, the between-subjects variation, is complicated by the fact that the variance across people includes a contribution to Tau-squared—another little technical problem, but, you know, both of these things, with a little bit of work and relatively

standard ideas, are easy to get around.

One caveat here is that we can only estimate Tau-squared, the between-persons component of variance, if there are replications in a study. If not, there's no information about between-person variability. If you only got one person, you can't get there.

But one of the things I learned from Rob and others who know more about this area than I do is that very many single-subject, single-case research designs involve replications and often several replications.

I'm only putting these formulas up there to give you a sense that it's possible to figure them out, not that you should understand them in detail and go away, you know, and be able to do this. But let me just say that if we assume that the structure of the autocorrelations is first order and the autocorrelation is Rho, the phases have the same length, and the variance is Sigma-squared, then the estimate of the within-person variance is this, and note that this—the formula

involves the length of the phase, the autocorrelation, how many people, and this quantity that I've labeled Si-squared, which is the computed observed variance within the baseline phase of the ith individual.

The idea is you don't get to an estimate of within-person variance directly by taking the variance of observations within persons, but you can get there, you can back it out with a little work.

Then, similarly, the estimate of between-individual variance, between-people variance, you can get from a quantity that I'm calling S-dot-B-squared, which is essentially the variance of the observations—the averages across people—but that doesn't quite estimate Tau-squared because there's this Sigma-squared component, but we already estimated Sigma-squared up there, so we can just plug the answer from up there down into the Sigma-half-squared here, and, in principle, this gives us everything we need to estimate the effect-size—the standardized mean difference effect-size—from a

single-subject design.

If you put it all together, you get something kind of horrible looking here, but it's, you know, computers do this stuff. You don't have to know the—you don't even have to know these formulas if somebody writes the program for you.

I'll make—and there's an expression for the variance of that quantity. So, these are the things you could use to, say, set up an estimate and a standard error or confidence interval, whatever. I'll note one thing, that there is a bias—this term 4m minus 8 over 4m minus five is a bias correction, which is important in a case where there aren't very many subjects within a study.

But the point here is that this, at least I think that this work is kind of an existence proof that you can produce a d-index that estimates the same thing as you would estimate from a between-subjects design but do it with single-subject design data, but there are some caveats.

One of the things that all the mathematics I did requires is that the time series is

stationary within phases. If it isn't, differencing might be necessary to make it more stationary.

I only looked at the problem involving first-order autocorrelations, but I think that's a reasonable first approximation.

A third thing is more troublesome. You need, to use this work, you need to have a value of the first-order autocorrelation that you believe. That's difficult. There's not much information in the data from one of these studies to estimate the autocorrelation. You can only estimate it with such huge uncertainty that you can't really rely on the empirical estimates from a single study.

But one of the other parts of our project is actually getting large—we're trying to get as much empirical evidence on autocorrelations as we can because, collectively, that gives us some pretty good information that you might use for imputing value to the autocorrelation for sensitivity analyses.

In some of the computations we've done, it actually turns out to matter less than you might

think within the range that seems plausible given the data we've seen.

You need replications to estimate Tau-squared. I think I mentioned that. Some of what I've done—but only at the end when you start producing confidence intervals—requires normality. None of the rest of what I did is—you know, I said, suppose these things are normally distributed, but it turns out you don't need them for the variance computations. You don't need that for the variance computations.

And I did mention here we're currently looking into the empirical values of autocorrelations, and we're surprised that they tend to be small on the average.

In the future, we're going to do several things, we hope. One is to extend these methods to the designs that will meet evidence standards because they'll have multiple treatment and baseline periods. That actually is reasonably straightforward.

We are developing empirical information

about the size of the autocorrelations, the size of the phases, and the typical numbers of study of individuals within each study.

We're going to be checking the small sample accuracy of the approximations that we're going to use to the distribution of d for, say, producing confidence intervals. And that's a bigger issue here than it is in other settings because the number of individuals is so small.

We're also checking robustness of sampling distributions to violations of assumptions about the value of the autocorrelation and about, you know, perhaps what happens if we have a little bit of second-order autocorrelation? Does everything sort of all fall apart?

Finally, we're working on the development of software to implement these methods, and the obvious reason for that is the formulas are a little bit complicated when you have balance. When you don't have balance, they get even more complicated.

And with that, I'll stop. Thanks.

[Applause.]

DR. BUCKLEY: Thank you, Larry.

I would like to now introduce Rob Horner as our discussant. You heard Rob's name mentioned in Tom's presentation. You heard his name mentioned in Larry's presentation. So, it's clear that Rob has made significant contributions to the use of single-case research in education.

He's the Alumni-Knight Endowed Professor of Special Education at the University of Oregon, where he directs the Educational and Community Supports Research Unit. His research has focused on developing evidence-based interventions that result in socially significant changes for people with and without disabilities.

In recognition of his achievements, he has received multiple awards—research and service awards—and he's worked for the past 15 years with George Sugai in the development and implementation of the School-Wide Positive Behavior Support, which is in over 11,000 schools nationally now.

Research evaluation and technical

assistance outcomes from those efforts indicate that investing in the development of a positive social culture is associated with improved behavior on academic gains for students, and, again, he has done an incredible amount of work in the area of single-case research.

Welcome, Rob.

[Applause.]

DR. HORNER: Jackie, thank you.

The role of a discussant is always an interesting one, and my view of the way that a discussant ought to operate is to really come back to what are some common themes that cut across, and what are some important implications for the work—but, before I get into that, how many of you have actually published a single-case study?

[Show of hands.]

DR. HORNER: 74.3 percent.

[Laughter.]

DR. HORNER: Okay. Now, so, in part—

DR. KRATOCHWILL: That's visual analysis.

DR. HORNER: Tom says that's visual

analysis.

[Laughter.]

DR. HORNER: Two big themes that I would really like you to take away from today. One, which I think happened right at the very beginning and happened pretty quickly, everyone should write this down. I want you to note that when Jackie stood up, she said, paraphrased, "IES considers single-case a valid and fundable research approach," and we really think so.

[Applause.]

DR. HORNER: That's videotaped. Right? Now, the second main thing, the second major theme that I think I want you to take away, those of you who have been doing single-case research for a long time, we are in an era right now—we've got about five, maybe seven years. This is really a point in time where we're changing what single-case research is able to do for the field.

We, in essence, as single-case researchers, are coming of age or coming out of the closet or moving beyond behavior analysis. But the

real issue is—and I think Larry and Tom both hit on this—we will not have the knowledge that single-case research delivers available to the larger education and psychological communities unless we gear up in terms of defining with better clarity and precision the kind of work that we do, the way that we do it, and how it gets interpreted.

The first thing I'd like to do is to really congratulate both Tom and Larry. Tom Kratochwill basically has been one of the champions, definers, and leaders in single-case research for over 35 years. We have all looked to Tom to bring us back into focus on things. Even though you listen to him, he has not always been random. He has been consistent, he's been clear, and he's been strident in terms of encouraging us to be precise.

Another key theme that you think is, is Tom was very excited when Larry agreed to be part of this symposium because, as Tom said, "Larry can both add and multiply."

[Laughter.]

DR. HORNER: And we do everything we can to juxtapose ourselves with Larry at different points in time.

But, seriously, part of what you're really seeing right now is a shift in terms of the way that we think about single-case research and come first to what Tom was talking about in terms of both standards and the use of randomization tests. We need to do a much better job of standardizing the criteria for effective single-case research if we're going to move beyond the small community of people who do this over and over and over.

I want you to think, for example, to what extent are we clear about what an appropriate training criteria looks like for training new single-case researchers? What are the criteria not just for conducting research but for accepting research for publication, for reviewing research for grant applications, and, especially in this context, the use of single-case research to identify and define evidence-based practices.

Now, Sam Odom led a committee that was

really focused on how we within special education and education generally look at defining evidence-based practice from our research methods.

The What Works Clearinghouse is really trying to take many of the messages that came from that effort and from the efforts that you're seeing up here and translate that into formal strategies that we can actually use.

We recognize that CEC, APA, NIH, the Autism Society of America, all have been engaged in the process of trying to say what are the standards that we use and how do we use single-case research to define things as evidence-based practice?

In part, the fact that we have so many people trying so hard is a clear example of the need that exists in the field. Everybody here—everybody here is engaged in determining right now what the standards are going to be because, as you know, these kinds of standards are not dictated by a paper. They're dictated by the convention. They're dictated by the extent to which we as a community agree that this is the way that we'll go

forward and determine what works and what doesn't.

I want you to get excited. I mean, as single-case researchers, seriously, a little heartbeat increase? Okay. Excited about being at the point in time when we're really watching the extent to which single-case research is going to move forward in terms of having a real contribution.

What Larry is bringing in his discussion of the d-estimator, I mean, in part, is really framing this logic that says this is possible. I know all of you were writing feverishly those formulas, and when he got to the delta-dot, I could see Charlie Greenwood's eyes just light right up.

[Laughter.]

DR. HORNER: But don't get stipulated on that. There isn't going to be one strategy that is going to run the gamut. All right. What we're going to need are multiple strategies. The critical thing is getting a better handle on exactly how are we going to make single-case research work.

The themes that I really am interested in

you coming away with: (1) the contributions to single-case research. We really need to come back to the issue that if we as a field are interested in people from low-incidence populations, we're going to need single-case research.

If we're really interested in conducting fine-grained longitudinal studies, the critical thing is, and one of the things I always worry about with the statistical models, is we're not talking just about a bunch of data points that were within intervention. We're talking about a group of data points that happened across time, and the change from time one to time two to time three to time four is critical.

We're interested not just in the basic pattern, but we're absolutely fascinated by the outliers. We know that Jeremy had a bad day, and we actually want to understand better what happened. It's that level of data that's going to serve as the core pilot information, that's going to much, much better guide some of our RCTs.

The other thing that I want you to really

come away with is the value of very fine-grained analysis—analysis across time that is looking very, very detailed at rates of acquisition in reading skills, at decreases of self-injurious behavior, at development of social relationships among young children.

The change across time and those small changes across time and how they add up is a critical thing that we don't want to miss within single-case research. But here's something I really want you to take away, and this is something that cuts across and is embedded in both of the presentations, from now on—whew, okay—from now on, when you're looking at doing analysis of single-case research, I want us to shift from the old way of doing it, right?

You look at the article and you say, "Does it document a functional relation?" And we're coming up with this standard of saying, "Do you get three demonstrations of the effect with those three demonstrations happening, each at a different point in time," right?

Instead, part of what Tom laid out and what is embedded in what Larry is talking about—remember, what Larry did was talk about two-phase comparisons. Take those formulas. Remember, the minor technical—that we always keep running in—I keep hearing that over and over again when I talk to statisticians. Well, we can do this, and there is this minor technical problem.

Okay. Here's part of what I want you to think about. From now on, teach your students this, (a) discrimination. If you're actually teaching a graduate course in single-case research, thing number one, can they discriminate designs that document functional relations from designs that don't? AB designs do not. ABAB designs do not. Single-case multiple-baseline designs with only two series do not.

You need three opportunities to demonstrate the effect each at a different point in time. In part, start with, regardless of the data, always look at the design. One of the things we teach, you never look at the data first when you

look at a single-case design. You always look at the parameter. What's the time variable? What's the dependent variable? And how is the design laid out?

Question number one: Does this design even allow the opportunity to demonstrate a functional relation? If the answer is "no," go get coffee. I mean, that's what's we do in the Northwest, right?

[Laughter.]

DR. HORNER: Okay. Second, only if the design has the power of being able to demonstrate an effect, then ask the question: do the data—do the data document a functional relation? Now, the thing that is a big challenge for us is in most cases when you get into the statistical models, you're always assuming that you're looking at level changes, but, in fact, and when people look at level changes, they look at level change in variability.

But notice what Tom did. When we teach visual analysis, we want to say you look first at the baseline, second at each phase, and then third at adjacent phases. You look at each case at the

level, trend, variability.

When you compare adjacent phases, you look at overlap, immediacy of effect, and the consistency of the pattern across similar phases. When you got two baselines, are the data in the baselines similar when you have to enter the B phases? Are the B phases similar?

Notice that there is no statistical model that we've seen at this point that actually takes all of that into account. Level, trend, and variability, got it knocked. All right. But, in terms of looking at the full range of variation, we struggle. Only—think about this—only if visual analysis says there's a functional relation, then I want you to—and only after you've said, "Yes, there's a functional relation"—then ask the question, "What's the size of the effect? What's the magnitude of the effect?"

And, at that point, then ask, "Is it socially important, and what conceptual contributions does it make?" Consider this, essentially, when you're doing both analysis of

dissertations, analysis of studies for publication, but, in part—here's what I want to say—as single-case researchers, we have been relaxed in the way that we go through this process.

We typically go straight to the core thing. Is there a functional relation? We haven't talked about the magnitude of the functional relation. We haven't talked about the extent to which it has all of the core features. We go straight to one place rather than talking about a continuum of an effect.

In terms of improving clarity with which we do this, one of the messages that I would take away—regardless of whether we move into some of the statistical pieces—I believe that one of the things that IES has done an excellent job of promulgating is this notion that even if you're really focusing only on visual analysis, we have not done this with a level of precision that warrants the impact that we want to have.

We've got to do a better job of training our students. We've got to do a better job of

documenting what we do. And both Larry and Tom emphasized that when you look at a single-case design, we've got to get over this AB effect stuff.

Most of the research that documents problems with visual analysis—documents problems with inter-observer agreement and visual analysis—is comparing AB effects.

We've been doing similar analyses when we look at complete studies, and we say, "If you look at the whole study, is there agreement on whether there is a functional relation?" The inter-observer agreement is dramatically higher, so part of what we're looking for, we've got to move beyond. This is not a single-case design.

All right. This is not a single-case design, and part of what we want is we want designs that allow multiple demonstrations of the effect. We need to do a better job of teaching people to look not just at the level but the extent to which the data within one phase project—right—growth curve modeling into the next phase.

Look at the confidence intervals around

the projected performance. Compare that with what you actually see. What's the magnitude of the effect? Level, trend, variability, overlap, immediacy of effect? Do that not just with the A and B effect, but look at now the reversal going from implementation back to baseline.

One of the things that I would argue is that going from baseline to intervention has one level of difficulty. Going from that first level—that first intervention—back to baseline is not exactly the same. Nobody has the weight that you would add to going back into reversal. We have not done our homework in terms of setting some of those models up.

Part of what we're learning, we're learning a lot about how to move from different ways—let me see if I can make this one. Nope. I didn't think it would work. I was going to show—one of the things I want to show is a way in which we actually teach those discriminations.

Some implications. One of the implications is—as you listened—you listened to Tom talk about

randomization. You listened to Larry talk about a d-estimator, and the d-estimator really was incorporating multiple statistical features. If we look at what's happening across the IES grants that have been funded, some people are using HLM models. They're using HLM as a strategy to be able to document the magnitude of the effect.

Swami and Breda and Dan and others at the University of Connecticut are using a generalized least squares approach, but it has many of the same logical setups in terms of looking at the extent to which we can actually parcel out the autocorrelation. Look at the extent to which what we get is different than would have been predicted—and to put that into a standardized format. To the extent that we can do that, we're going to become much better.

One of the things that we did recently at the University of Oregon is we had a seminar with advanced graduate students looking at advanced analysis of single-case research, and we actually went through the literature on each of the

different methods.

One of the things we came away with—and Gina was actually kind enough to be one of the leaders of that seminar—was that we as single-case researchers need to do a better job of looking at how we actually define our research questions.

Typically, we say, "Is there a functional relation between the independent variable and the dependent variable," right? If we really want to use HLM, and we want to look at some of the Beta weights that are embedded within that, we need to do a better job of saying, "Are we really interested in change in trend? Are we interested in change in level? Are we interested in change in variability?"

For example, here's the simulated study. ABAB study. No real change in level. No real change in trend. Is there a functional relation? The way in which you define that would be based on the way in which you build the research question. Some of the things that we're learning, we've got to get a little more precise in terms of the way that we

actually build our designs, build our questions.

Some other things that we're learning. All of the pieces. I mean, I love the fact when Larry and when Tom were talking about the statistical pieces, you start getting into, well, you know, how many data points per phase and is there an equal number of data points in baseline and intervention? And the 73.4 percent of you who raised your hands, you all know, there were not an equal number of data points in baseline as there were in intervention, right?

And you know that some of the time when you do those reversals—especially, you're doing a reversal, say, with a kid who's engaging in a mild inappropriate behavior, right? There's not a large advocacy for maintaining the reversal an extended period of time, right?

[Laughter.]

DR. HORNER: In part, what is the rule? Part of what we're coming up with over and over and over, if you have less than five data points per phase, it's going to create all sorts of problems.

So, if you want to do—right—and it's better to have 10, but if you're getting down to less than that, we got a problem.

Randomization is going to create all sorts of opportunities. But many of you were also going through the logistics of how you would pull that off. Part of what I want you to come back to, this is not about defining the new way of doing single-case research, it's about expanding the array of options that we have, and part of what I think Tom is putting on the plate is, let's talk about what would be the conditions in which you would use randomization and make it work.

The big implication. Single-case research is an effective method for documenting causal relations. All right? This is a valid way of actually doing science. Part of what's got to happen, though, is if we're going to use those methods to really deal not just with basic questions but things essentially—some of those questions that are really hard to do get at with group designs—we're going to need more than just

the traditional single-case issues.

We're very interested in meta-analysis. We are very interested in being able to say, "What does the body of literature say?" Now, I got to tell you, honest truth, I think that for a group of single-case researchers, we do not need a statistical model to say this is when we would be willing to consider these single-case studies to document an evidence-based practice.

I think we could come up with a standard, and, in fact, Sam led a project in which we did exactly that. The problem is we're building a standard that only we will acknowledge. If we want the knowledge that we build to actually have an influence on the larger psychological, educational, and intervention community—the community for prevention science, for example—we must speak in the language. We must demonstrate that we've got the tools that make it happen.

Meta-analysis is that tool. And unless we can reach that bar, we'll be forever limited. So, part of what I'm very excited about—and what I

think is really an important message—this session is really about setting a trajectory in which we both honor everything that single-case research has been.

This is not about changing single-case research. This is about, one, doing one heck of a lot better job of teaching single-case studies and doing a better job of defining our technology with a level of collective agreement that we can actually be part of the larger community.

I think that in terms of using single-case research, we want to be able to understand the behavior of individuals much better than we have in the past. And there are things that single-case research can do that no other design option allows.

Second, we want to really be able to focus on change across time, not just across two points in time. We want to be able to talk about cumulative points in time, and I know those of you into growth curve modeling, our eyes light up, but this is something that's important. The multitiered interpretation—design, visual analysis, statistical

analysis—are going to be things that you can use that are going to be important ways of making that work.

Now, think just for a second in terms of doing visual analysis—in terms of teaching visual analysis—instead of doing what we typically do over and over again, which is only show studies that work, I want you to think about doing something like this. So, here you are. You're teaching single-case research, and this is a multiple-baseline design. You're going to teach never look at the intervention first. You always look at the parameter. Is this a study? Is this a design that would allow documentation of a functional relation? Look at the baseline data—level, trend, variability.

Look at the intervention—level, trend, variability. Compare adjacent phases, overlap, immediacy of effect, consistency. Look at the extent to which when there's change in one series, there is no change in the others. Then, on a scale of zero to seven, do these data document a

functional relation? Got it? All right. Now, do it again.

Got it. Zero to seven. Zero to seven. Zero to seven. Zero to seven. Zero to seven. Look at subject three. All right? Think about a situation in which your students can do this at a rate and with an accuracy. If you can get .95 correlation with Tom Kratochwill's visual interpretation, you, too, can read single-case research.

[Laughter.]

DR. HORNER: For the What Works Clearinghouse. Thank you.

[Applause.]

DR. BUCKLEY: Thanks, Rob.

We are running really short on time, but we could have time for maybe one or two questions if someone wants to use the microphones, please, so it can be recorded.

DR. GOLDSTEIN: Howard Goldstein from Ohio State University.

One of the things that I don't hear get talked about very much is the context for looking

at effect-size metrics, where we talk about conventions that we use for educationally relevant measures, and one of the things that concerns me a little bit is that one of our challenges in getting an effect-size measure, a d-estimator, or whatever that's going to be comparable for single-subject design and group designs is that, oftentimes, our measures are quite different.

And I have the suspicion from some of my own work and people that have been doing work in this area that our effect-sizes look much larger than group design effect-sizes, and I'm wondering whether that's a fundamental difference that relates to kinds of measures that we're using, and if there are going to be other strategies that we may need to think about adding to the way in which we conduct single-subject designs as well?

DR. HORNER: I'll start in responding. Howard, you asked similar questions three years ago when we did this, and I think you were right there, and you're right now.

DR. GOLDSTEIN: I don't remember.

DR. HORNER: Yeah.

[Laughter.]

DR. HORNER: Yes, I think measures always make a difference in terms of effect-size, and I think one of the reasons why single-case studies are going to have very large effect-sizes is because most of the measures are very, very fine-grained measures. They're very sensitive measures.

When you use dull measures, you have much greater error variance, and you have lower effect-size.

I think Larry's comment about the autocorrelation, I think, also is going to be important. We've always assumed that the autocorrelation was very large. We also have calculated autocorrelations. We've actually found some that were very small and some that were substantial, and we weren't necessarily able to predict as well as we thought we did.

We did not do a good job in visual analysis in terms of looking at that. The issue, I think, is also we've got to be careful. In group

designs, they do this odd thing where they actually keep the people who don't respond, and they publish those data, and, in single-case research, when people don't respond, and there's not a functional relation, we oftentimes don't see that published.

We've got a publication bias that is going to be a major concern when we look at comparing the effect-sizes in meta-analyses.

I think we've got issues that we need to determine, but the work that we're doing first, I think, is coming up with something that we can start with being comparable, and then I think it's going to be great fun. It's going to be a minor technical complication we need to do.

[Laughter.]

DR. HEDGES: Minor technical, you know, problems are the way I make my living so—

[Laughter.]

DR. HEDGES: I would just add to what Rob said—that the phenomenon of effect-sizes being a function of the outcome is true in between-subject designs as well. It's very highly, frequently

replicated finding that the more aligned the outcome measure is to the treatment, the larger the effect-sizes are, and some people get disturbed about that, but, on the other hand, when you measure an outcome that the treatment really doesn't have very much to do with, it's not surprising you don't get very big effect-sizes.

Ideally, if you have a treatment that does work, and you measure something the treatment is supposed to do, maybe we have to get used to big effect-sizes.

DR. KRATOCHWILL: I guess—oh, you want to take another question? I had a couple of comments on that just that I would add. You know, I agree with what Rob and Larry said—that the measures are different. The fundamental measurement systems are often different, you know, if you look at, you know, the work that Johnston and Pennypacker introduced in terms of the different worlds of measurement, it is totally different, and conceptually, there are different issues there with the types of measures, repeated assessment versus

pre/post and so on.

A lot of the measures that are invoked have normative standards that are validated through traditional psychometrics, and that doesn't necessarily invoke a social validity context. If you get excited about a half a standard deviation change on this, you know, the rating scale that you've used, that's very different than reducing aggression, you know, by 99 percent, for example, so the measures are different.

But I think a lot of the work in single-case design has been guided by social validity constructs. I mean, we've all looked at that and talked about the meaningfulness of the effects. So, that has, in some ways, it's been a positive bias.

The other thing I would say, though, in some ways, I'm not sure we know the difference in the effects because some of the calculations that have been used to determine effect-size in single-case designs have been problematic.

For years, for example, I was depending on some of the work that Busk and Serlin did, you

know, which was a chapter in our '92 single-case design and analysis book. There are problems with that, and so, when we find an effect-size of 15.0, it doesn't translate very well, and so, I think one of the things, hopefully, that Larry and David and Will will do is tell us about the equivalency and the differences in those kinds of measures that are used across the two types of research. I don't think we have that information really very well right now.

DR. BUCKLEY: Stephanie.

DR. PETERSON: Hi. I'm Stephanie Peterson from Western Michigan University, and I just have a comment and a question.

My comment is, "Thank you" to you guys because I think this is really important work. When I was listening to the talks this morning about the charter schools and how they were asking questions about why does this intervention work, you know, what is the particular thing this teacher is doing, all I could keep thinking is single-subject designs could answer those questions in a very elegant way.

I would, I hope that we'll see those researchers starting to adopt some of our design methodologies as well.

My question is hopefully not naive and silly. I'm always apologizing to Tim Keith, who is doing the statistical analyses on our data because I tell him, "When you explain it to me, I get it," just like today, I get it—what you're saying. Then, I walk out and I try to explain it to somebody else, and I realize I really maybe don't get it.

[Laughter.]

But one of my questions is, so, for example, if we have an ABAB design—and Rob, you sort of alluded to this—we could have, you know, an upward trend, say, in our A phase, a downward in B—up and down—and if we look at the means, they're all the same, but the trends are very different, which causes some problems when we're thinking about analyzing the means of our data.

I think about the work that our colleagues do in experimental and basic research where, say, if they're working with rats and pigeons, they'll

first go through a learning phase with the rats, and then they only do their statistical analyses once they've reached a stage of stability, and so, they'll conduct their statistical analyses on that subset of data.

I just wonder if any of you have thought about that and how that would fit into what we need to do here? I'm interested in your comments on that.

DR. HORNER: Well, I'll start. First off, Sidman told us in 1961 that we should do analysis focused on steady states versus transition states, and part of the reason is waiting 'til you get to a steady state until you're able to say this is what the true effect of the intervention is.

The study that you mentioned, which is, you've got a downward trend, upward trend, downward trend, upward trend, if all you're interested in is demonstrating change in trend, that's going to be cool.

If you really are and if your research question said that you're interested in level

change, then clearly you intervened too early, right? If you've got the—you're going to let the trend go, that's an example, and, actually, it's one of the very, very, very common mistakes that young researchers do, is they intervene because they wrote in their proposal that they were going to intervene after seven days.

And you just, you know, until you're using a randomization format, you've got to stick with it based on actually looking at what the data set goes.

I think the thing that is really important—first off, in terms of interpreting the data, the data, I believe, can be interpreted if what you say is, "I'm really interested in showing change in trend." If what you said is, "I'm really interested in looking at change in level," then there were errors made in the analysis—in the way in which you developed the plan.

That's why I think a really big issue and one of the things that we really, as single-subject researchers, need to become much better at is

defining with precision the research question so that we're not just saying, "Is there change," but, "What is the dimension of the behavior upon which we expect change?" And when we do that, I think we're going to be much better set up to work collaboratively with our statistical partners.

DR. HEDGES: It's hard. I think you covered the water there. I don't know if there's much to add.

DR. PETERSON: Well, I guess my question then is in statistical models. Should that first part of the phase be included or should it only be the part where stability has been reached? I guess that's—

DR. HEDGES: Well, I guess I would say this, that, you know, Rob is exactly right. You need to be clear first on what it is you're trying to affect, but I'll toss out one other thing. One little piece of the work that I do, have done, is experimental cognitive psychology, and I'm used to sort of working on experiments and trying to get them to work and then not paying much attention to

all that pilot stuff and then doing a real data collection. This is a run-in period in which you try to figure out how to do the experiment.

But then, typically, you can get the same result over and over again if you run the experiment. I don't think it's, I think it should be thought about to some degree, in some cases, as a learning experience for the researcher, just figuring out how to do the study right, and then, ideally, if you do it once, you can replicate it as many times as you want. For what it's worth, that's the way I think experimental psychologists think about it.

DR. HORNER: The one quick—in terms of the detail—if what you're looking at is of immediacy of effect is what you're looking at, then using the last five data points of baseline/the first five data points of intervention is perfectly reasonable.

If what you're looking at is level change, then you should use all the data within the phase. That would be my—

DR. BUCKLEY: Yeah, we're a little bit over time, so I'm sure—and I know you've been standing and have a question—I'm sure they wouldn't mind hanging around for a few minutes to answer your question, but we really are over time. So, I want to thank our panel one more time.

[Applause.]

[Whereupon, at 4:38 p.m., the panel session was concluded.]