

Service-Learning Evaluation: An Overview

Presenter: Shelley Billig

>> In training projects from the Corporation for National and Community Service. That's the parent project of Learn and Serve America, National Service-Learning Clearinghouse. We're recording today's webinar so while we will be muting your phones, later on we may open up for Q&A. So if you will be participating in Q&A please keep in mind that this is being recorded. If you're unwilling to be recorded then please don't ask a question. Larry do you want to go ahead and start recording?

[Background discussion]

>> Hi everyone. This is Liberty Smith, Associate Director of National Service Resources and Training. This is the parent project of Learning and Serve America's National Service-Learning Clearinghouse. On behalf of SOSA Clearinghouse, Learn and Serve America, and the International Association for Research on Service Learning and Community Engagement, I'm delighted to welcome you today --

[Background discussion]

To our co-sponsored webinar, Service-Learning Evaluation and Overview with Shelley Billig from RMC Research.

>> The conference has been muted. [Background sound effects]

>> So once again we are recording today's webinar. That means it will be available from the National Service-Learning Clearinghouse website. But please do keep in mind that if you un-mute yourself in the future during Q&A you are being recorded. If you're uncomfortable with that please don't un-mute yourself. Before turning this over to Shelley Billig from RMC Research I'm going to pass this off to Scott Richardson from Learn and Serve America to just tell us a little bit about the importance of evaluation to learn and serve in the rest of the field. Thanks.

>> Terrific, thanks Liberty and thanks to Shelley for being willing to spend an hour with the service-learning world, the folks who are interested in the evaluation question that we have before us today. As folks know government sponsored programs in an era of tighter budgets need to demonstrate evidence that they actually have the effect that they claim to have. And in Learn and Serve we always talk about how does this affect the young people -- the participants and students who are involved. How does it affect communities and improve conditions there? And how does it affect institutions and help root service-learning in community organizations, higher ed institutions, and K-12 schools? And to help us again to look at the best ways to go about gathering evidence that service-learning works. That it's an effective way to come at education and community and civic engagement, we've asked Shelley to talk about some of the strategies of laying out purposes, questions, instruments, and reporting when it comes to evaluation. Because we here at the national level, we're always asked by Congress and other stakeholders, okay show us the evidence. And that conversation is happening on parallel tracks at the state level and in communities all over the country. And so we got the prolific researcher of service-learning and education, Shelley Billig, who will take it from here and get us going. Thanks again to Shelley. It's all yours.

>> Thanks Scott and hi everybody. Please -- I think you probably heard Liberty say that we're going to try and hold questions back, but if you have a burning question that you want or something that you need to have clarified, if you'd just type it into the Q&A section down on the bottom and send it we'll try and get to it as soon as we possibly can. As usual, I've got too much information for the amount of time that I've been allocated so I'm going to rip through several things in the hopes that we'll have more time

for discussion and Q&A at the backend. So you see on the front page how to get in contact with me via email if you have additional question or want to talk about something specific to your program. This is a general overview as opposed to a [Inaudible] session. So those of you who are familiar with evaluation you'll see a lot of things that you already recognize. Here's what I thought we would talk about. First, we want to look at why you should evaluate and Scott's already given us a heads up on a lot of good reasons. I want to briefly talk about the characteristics of effective evaluation. And then start getting into the technical pieces. What a good evaluation question looks like. The need and use of a logic model as a guide. The kind of evaluation designed to promote as being more rigorous and effective than others. We want to look at methods you can use. I'm going to talk about and present to you a few sample survey subscales. The survey subscales I'm going to present are things are related to some of the major questions that we have in the field. I'm going to talk about sampling, even subjects protection. We'll do just a very brief piece on analysis, drawing conclusions, and elements of a quality report. I'll talk a little bit more about using results for improvement and give you some resources. As I go through this, just to try to make it a little more concrete for all of you, I'm going to talk about the Learn and Serve America Cluster Evaluation that we're doing. Cluster Evaluation currently involves about ten states and a couple of national projects. And what the Cluster Evaluation folks, partners of them have agreed to use some common core areas of measurement and method and then some unique aspects. So I'll try to use those to illustrate the kinds of things I'm talking about. I'm not going to give you the names of everybody in the cluster, but I do want to express my appreciation to them for allowing me to talk about them. So quickly if you would just think about and individually write down the most important reason that you think that LSA should evaluate it's programs. I'll give you all of 20 seconds to do that.

^M00:06:48 [Silence] ^M00:06:56

>> And here we go. Common reasons that are out there are to document your outcome, to see what your objectives were met, to improve your program, to procure additional funding, which a lot of people say because it's required. And Scott mentioned what we're finding with federal and state funds is that the accountability pressures have increased dramatically. And with those increases in accountability people are not willing to accept some of the evaluations that we have done in the past, simply because they're anecdotal in nature or they're not rigorous enough for the results to be reliable and valid. And so with the push for more accountability and particularly in times where budgetary constraints are rearing their ugly heads, people only want to fund the sorts of things that are most effective. And so we want to make sure that we do the very best job we can with the funds available to make our case that service-learning is a fabulous and wonderful and extremely effective way for our young people to reach a lot of desired outcomes. What we want to make sure we've got in these evaluations is first that they are accurate. As I'll be talking later on in some detail, we can too in our field do a little pre-post evaluation. The problem with those is that while they see a difference every time, really if it's just a pre-post you can't contribute it to service-learning since you don't know what else in the world could have influenced the outcome. Maybe they had a better teacher. Maybe they had a good day. Maybe they had other events in their lives. So when you just do a pre-post evaluation basically people have ripped it apart and said that's not anything that will count towards the current evaluation. It's a great element and they have lots of exploratory information, but they are good for hypothesis generation, rather than hypothesis testing. And in this era of accountability there's far more emphasis on impact than there is on exploration. And so you have to be careful if you're doing just an exploratory kind of study because it can't be used for some of the purposes that we would like to see. It is very good for other purposes and we'll discuss that a little bit later. We need to make sure that what you're evaluating is pertinent. We often find that people go off on a tangent and don't actually answer the evaluation question. Evaluation questions tend to be about impact and implementation and factors that influence outcome. And

sometimes people will go off for several pages without even explaining the answers to those questions. They need to be objective. We basically want us not to [Inaudible] opinion to the extent that we can't. When you do do that you need to label it and talk about it in the discussion. But in general you want to faithfully represent your data and then label your interpretation as such. Things need to be well organized and readable. We as a research company and several others out there can be a little jargon-y and very technical. And while that suits the purpose of certain audiences, it really doesn't suit the purposes of others. And so we want to make sure, especially for communicating with public or several of our tons of stakeholders [Inaudible], that we have a communication that's relatively jargon free and easy to understand. It's needs to be logical. This one sounds like it's a no brainer, but unfortunately many of the critiques that we do we can't see a line from one thing and another. People are drawing conclusions that are wildly uncalled for. And so we really got to make your case by providing whatever evidence for your findings, in a way that people believe is valid and reliable. And finally it needs to be useful. Investigating unnecessary questions in a world where finances are limited is not always the greatest idea. And so we want to make sure that we have information for improvement. The sample as we said we got the Learn and Serve programs in multiple states. All of these are state evaluations with the exception of two wonderful national programs, that we're looking at. They're always the same core evaluation questions, design, survey subscales and analytic framework, and they have a limited number of customize evaluation questions and survey subscales and analysis associated with that. Those of you that are finding funding to be a challenge may want to consider pooling your money in much the same way that our cluster pooled their funds. Because cluster evaluation is far more efficient, it's much less expensive, and it allows for cross state accreditation. And so it's got a lot of advantages to it. And so if you're willing to kind of throw in with people who are doing projects similar to yours, there's huge advantages to doing that. Most evaluation questions that we're going to ask are going to be impact evaluation questions because often times what our stakeholders want to know. And in particular the big key questions for our field is what is the impact of participation in service-learning on the youth and the adult person. As Scott mentioned earlier, our key constituent here is the young person who is participating, but it's also the community. And so the idea is if you're going to be engaged in service-learning you should be making a community impact and you should be able to measure that. And so the areas of impact should be clearly classified in advance, your answers should align to it, and hopefully if all goes well you'll be able to see positive impact for both the young people, the adult participants, and the community they're serving in measurable ways. Those are the keys. And if you go nowhere else and these are the two big questions you ask, you'll be making a huge contribution to our field. In our contract with the cluster, a lot of people have said well, you know, there's a lot of ways that you can impact young people, what is it you specifically like to see? The cluster is a school based program and so if you want to keep service-learning in schools, what people want to see are some academic outcomes, which makes a lot of sense. We in service-learning have logic models that I'll talk about in a moment. And a lot of our logic models say what service-learning really does is give students the opportunity to engage in meaningful and authentic work. By doing so it raises their academic engagement. And by becoming more interested -- persisting in their academic work and actually learning the content that's getting applied within the service-learning activities, people are doing better. And so we have deconstructed some of this to take a look at specific areas for young people who are in the school based service-learning. First, we're looking at academic engagement. When we talk about academic engagement here at RMC Research, we're defining it as affective, behavioral, and cognitive engagement and self-efficacy and confidence. What this means is affective is being more interested in subject matter in which you have lent to the curriculum. And behavioral we're talking about being more willing and persisting in those tasks that you're assigned to do -- the academic tasks. And cognitive engagement talks about the actual intellectual engagement with the content that you're learning. And the academic self-efficacy or confidence is something that's really important to us for two reasons. One is that it's

really good measure of academic engagement. Self-efficacy means you feel like you can do it, you have to ability to master the material, you're confident that you can do it, you have a sense of competence. Another thing we like about both of these measures, the affective behavior and cognitive engagement and the self-efficacy and confidence, is both of them are strong predictors of staying in school. And so we use them in our work both as academic engagement by itself and it's one of the several predictors we use for drop-out prevention, which I'll talk about in just a minute. Academic performance and achievement is what Congress wants to know about. Basically, they're interested to know, does participating in service-learning actually make a difference on your test scores. And so in order to take a look at that we are gathering test scores. The test scores we're gathering are the same state assessment scores that you use for any project because they want to make sure people don't think we have customized our measures -- our objective measures to our approaches. So we've been using English Language Art scores, Mathematic scores, Science scores, and -- where available Social Studies scores. Also where available, we're looking at writing scores. And then the other thing we're looking at as part of both academic engagement and achievement is attendance. And the idea here is to look at the attendance in service-learning classes, as opposed to non-service-learning classes. So if there's something about service learning classes that keep disengaged students in school because they really want to be a part of the service-learning project. And then we're also looking at disciplinary referrals because we find that young people who are more engaged tend to be less bored and cause fewer problems. We've had a fairly good literature review on this piece of it. We've got some nice measures on it and so that's something that RMC's looking at. Our third common area on our clusters is dropout prevention. So we're looking at the academic engagement measures and then for older students we're looking at aspirations for graduation, post secondary, and career aspiration and interests and intentions, enrollment in AP, enrollment in SAT, etc, and actual dropout rates. The other common areas acquisition and 21st century skills, several of our states, but not all of them are looking at this. You can see the way we have operation wise the 21st century skills. And we have used the concepts that came out of the framework for 21st century skills from the partnership for 21st century skills acquisition. We look at temp skills. There's a lot of interest right now to see the extent at which they're declining is helpful. And getting students to become interested in and master concepts in science, technology, engineering, and mathematics. Of course, we're looking at environmental stewardship as a lot of our programs deal with the environment and this is something that is very important to a lot of people. When we look at stewardship we compose that into knowledge, skills, and disposition. We also look at the aspiration to continue work in the field. A few other areas that are common and we look at as well, civic engagement, activities, responsibilities, school attachment, and community attachment. We tend to the constructs from the literature at large rather than the stuff that's been developed specifically for service-learning. We look social-emotional learning as you can see several of those concepts there. The difference between some of the 21st century skills and the social-emotional learning is that the social and emotional learning are individual level aspects, whereas the 21st century skills tend to be durational aspects. And so they have different variances you tend to get different kinds of responses to each of them. And so they are both in their own rights interesting to look at. Then finally, as you all know quality matters. Every study that has been done in the last decade or just about every study, has determined that high quality service-learning has some positive outcomes and no quality or low quality service-learning has no outcomes what so ever. And so we need to do a better job at understanding what quality is. We are using the K-12 service-learning standards and indicators of quality practice. We are taking a look at each one of them separately and as cluster to see whether they predict outcome. And just as a heads up they are holding together extremely well. As a group they completely predict outcomes and it's the difference between impact and no impact. And we're also finding that some of you may be because some of them have more weight than others. Another set of typical evaluation questions that are asked and should be asked are the things called mediators or moderators. Mediators

or moderators are our fastest approach to influence outcome, so that if you can control them then you can actually see differences in your impact based on the controls that you put in. They tend to fall into two kinds of categories. One of them is participant characteristics and so it might be the demographics of the adults or the students. Student achievement levels tend to predict outcome. Teacher experience and so forth. For example, we find that teachers -- and we've pretty much found this consistently -- teachers with more experience tend to have better outcome than teachers with less experience. The magic numbers seem to be somewhere around two or three years -- usually three years. And so because we know that it makes a big difference in the way we analyze our data. We're finding very interesting things -- differences between student groups based on their ethnicity and on their age. And so these are wonderful questions to explore. The second set of questions helps to deal with program design factors. We're particularly interested in the set up and quality of the programming, but we also know things like professional development can make a difference. And so the more training you get in service-learning, theoretically at least, the better you do. And we have some mixed data, at RMC, on that that I can talk about if you ever have some time. The kinds of things we look at for these you can see listed there -- gender, age, and experience, content expertise, social, academic status, and all kinds of other participant characteristics. And then the program design characteristics, besides the quality, one the things that's popping a lot is whether there is direct or indirect service. You know for direct service, you need to have direct contact with those being served. Indirect meaning you don't have contact with those being served. And that for us is starting to predict results. Other people and in the past RMC's looked at things like whether or not there's efficacy component in the service-learning program, which is also program design characteristic. And that seems to be making a difference difference. The type of demonstration and so forth. So there's lots of program design characteristics you can explore. Those of you who are listening, who are working on a presentation this is a great area that we would like to encourage you to pursue. We strongly recommend using logic models as a guide. The corporation requires logic models. Most of the logic models we've seen have not been as high quality or rigorous as we'd like to see them. It's worth spending your time on your logic model because it will dictate your evaluation questions. Logic models have inputs, outputs, outcomes, and factors that moderate outcomes. The difference here and I'll give you an example on -- wow, that's pretty small, but there's an example of a logic that we've used. You can see that the input has to do with funding and support and professional development and all kind of things. The outputs are the typical kinds of things that you count -- participant hours, numbers of people who participant, things along that -- participate, excuse me, things along that line. And then we have divided this into the tell all categories. And those are the short-term, medium-term, and the long-term outcomes. You can see here that the way we choose to do our logic model is to try to thread through the outcome areas. So when you see academic achievement it moves from short-term to long-term and through middle and it's got some similar measures and some that are longer term. I'm hoping if we get a chance to longitudinal studies we'll be able to do this -- to measure all the way through the long-term outcomes. You may be familiar, but since this is an overview let me quickly go through the major types of evaluation design. Experimental designs are still considered to be the gold standard. They're where you randomly assign students or classrooms for schools or districts or other kinds of units to treatment and control. The treatment would be service-learning. The control would be no service-learning. There's a lot of challenges with the experimental design that I'll go through in just a minute. The strongest challenge of which has to do with cost. It tends to be extremely expensive to do well. Quasi-experimental design, such as matched comparison groups, tend to be the most popular and reach the greatest figure without the expense of an experimental design. The matches are really important because what you need to do is to match students based on factors that you know influence the outcome areas that you have. For example, we know that current achievement level predicts future achievement level. And so if you're doing well you tend to do well in the next level. And so you have to control your current achievement level if you don't

match well. We know demographics matter. We know enrollment in gifted programs versus traditional programs versus alternative education programs or special education programs matter. And so the degree to which you can match your groups is incredibly important to the rigor of your evaluation. A lot of people use pre-post designs, as you see right now I'm not so much of a fan. You do what you can on these except that you can do it possibly better than nothing. And case studies which I'll talk about at some length. With experimental design it does allow attribution so if you possible can randomly select and assign people this is the best way to go. Because you can specifically control for things, you can measure all kinds of things, and it will allow you to say service-learning was the reason for any differences you might see, positive or negative. Random assignment can be at lots of different levels as you see. But the biggest challenge is it's really difficult to get people to agree to random assignment. A lot of times when you want to randomly assign students, for example, their parents get upset and want them to be assigned to one teacher or another or to service-learning or not service-learning -- they have opinions and as soon as you start to accommodate any of those opinions then you have lost your random assignment or you have to exclude those students from the study. The other issue with random assignment -- and one way to get around all this is you can do a staggered start which is what a lot of people do. With a staggered start that means that everybody is in a pool for treatment. Some people get the treatment the first year, some people get it the second year, some people get it the third year, and so you have natural occurring groups -- experimental and control groups and exposure groups. So it's one way to do this. With random assignment you have to sufficient sample sizes so you can detect a potential small affects. We think the affects of service-learning are quite small. We believe the affect sizes may be around .03 which is really difficult to detect. And so in order to detect that kind of size you really need a fairly large sample. With quasi-experimental designs, as I said it's the match that really matters and that's the most rigorous of all the designs that are out there. Once again, you've got a control for influences and you still need a sufficient sample size to be able to do the types of analysis that are appropriate for quasi-experimental designs. Pre-post, just before and after a program, it's much weaker. As I said earlier because you can't contribute the outcome to the intervention. Too many other possible explanations of how they've been eliminated in a pre-post design. And often times, I know we've done this and you may want to try it yourself -- we'll do something and just analyze our treatment group and then we analyze our treatment group with our comparison group and all of the significant differences disappear. And so if you take the same data and do this you'll see the flaws in your pre-post design. The case studies typically include in-depth qualitative investigation. We like case studies a lot when we're trying to either portray what something new looks like, when we're trying to explore something, if we want a more in-depth exploration of any given program design characteristic or something like that we use [Inaudible] and observations when we do case studies. It's really necessary you make sure and triangulate your data, which means that you have three sources that all measure the same thing so you can see whether there's consistency or not in the results that you're getting. Rigor is still important. There are standards for qualitative research that should be applied here when you're doing qualitative case studies. There are rigor standards for focus groups, for interviews, for observation, and for the way that you analyze and reduce you data. And so we're not saying that this is less important than a quantitative study, it's just different from a quantitative study. It's used for different purposes than a quantitative study, but it still needs to reach the highest level of rigor. The design again depends upon your question, your cost, your timeline, and everything. There's really strong advantages and disadvantages to each of them and we posted a handout -- I'm not sure where that went -- but that summarizes the designs. So I'm thinking that most of you are more familiar with this and can even look at that handout if you want more information. Our cluster uses a quasi-experimental design with matched comparison groups. We had to get into the situation where I think some of the rest of you are, where we had to start with the retrospective pre-post as a pilot and a baseline because the funding came so late in the year that people had already started the service-

learning before we could get our pretest done. If people have already started service-learning before the pretest you get a ceiling effect, which means that your impact won't be as high. Those of you who have worked in the field a long time, we tend to get a fairly impact at the very beginning of the service-learning because everybody gets so excited about it. And then there's an implementation dip where the excitement actually goes down when kids are doing the hard work of investigation and some of the other pieces that need to be done towards the front end of service-learning. And then we end to get it going way back up again as they plan and implement the service and have contact and start to see the difference that they're making. Demonstration is an up and down kind of process because a lot of students don't like the preparation for the demonstration, but then they love the results. And so depending on when you're measuring the impact of service-learning that actual timing can be important in looking at your outcomes. And so you want to make sure that your timing is good. That's if you do your pretest, of course, before service learning begins. And you do it all at the same time. You really shouldn't have more than a three week window in your sample of when you do your pretest and when you do your post test. Those of you working in schools know that you have to avoid things like test windows and so forth because all of those things can also affect your results. The kinds of methods that we're talking about using most often here in service-learning are surveys, a number of qualitative methods like focus groups, interviews and observations, the objective data that we talked about before like the test scores, essays can be objective or subjective depending upon what the prompts are, and your [Inaudible] reliability and how well you devise your rubrics and other scoring criteria, and several other things. Generally, when you're doing the survey subscales and we in the cluster are using surveys, you want to make sure whatever survey subscales you take are related to your logic models and impact areas. You have to make sure that they're actually valid and reliable. We have seen in our review of the literature an awful lot of people using subscales that have very low reliability. Those of you who remember your 101 classes, validity is that you're measuring exactly what you say you're going to measure. In other words, it is the construct. Reliability means that if you measure the same thing over time or in multiple ways you'll always get the same answer because you are actually measuring that construct and it's reliably done. Most people when they're judging a survey subscale use what's called Cronbach's alpha, which is the measure of internal reliability. Most journal articles report the alpha sign and you can say your criteria, but generally you want try and get an alpha that's at about .7 or .8 or higher. Otherwise, you need to conduct item analysis which is not nearly as good as using scales in terms of a rigor study. And so take a look at those alphas. You need to make sure your survey is actually coherent. A lot of kids get confused when they're taking surveys because they can't see the logic of the survey or things aren't labeled well or they, you know, the response sets are not the same and so it's sort of jarring to them. If you tend to use, for example, a five point response you should use that consistently and not go from three to five to four to seven because that's just confusing for your respondents. You need to make sure it's the appropriate length. Pretty much after 20 or 30 minutes everybody gets fatigued when they're taking surveys and they don't tend to read the questions carefully or give you accurate information. And again it has to at the readability level. Sometimes we find that the words that we use are too difficult for the young people who are taking our survey. And what is recommended generally is that if you're working with middle school students that you use a readability level that's at about fifth grade, if you're working with high school students you use an eighth grade readability level, and when you're working with adults you use an eighth grade readability level. And you can on your computer -- at least we do in reflections on some of the words and stuff you can actually run it through for readability and it'll tell you the grade level at which your text is found. I wanted to give you a few samples because this tends to be the thing that people ask us the most about. We at RMC have collected literally hundreds of samples in the areas -- the main areas that we're most interested in and we tend to keep the ones that we think match our designs the best. I want to give you a sample of just two things so that you can see there's lot of ways to measure these various things. The

one in front of you right now is our survey of community engagement. This is how we kind of organize the way we think about things. This particular scale is for grades 6-12. It's got both basic content validity, it's alphas range -- that's the internal reliability measures over .80 consistently. We've actually been using this scale for years. And so it's a really good alpha, very reliable. And you can see the prompts. What I wanted to point out in this prompt is that we're not making the assumption that people understand what a community is. We're actually defining it in the prompt and that's one of the flaws that we often see when we're looking at studies, is that people assume that when you say community everybody knows what you're talking about. Therefore, when you get your results honestly you don't know how to interpret them because you haven't given them a definition so that you know that everybody's responding in the same way. And so you can see that we define that in there. We have six items in this scale. This scale as you can see sort of ascends a little bit, it's got two different dimensions around community engagement. All of these came from a literature review that we did around what community engagement tends to look like and the various aspects of community engagement. And what we find here if we use an agreement scale. The only difference between this scale and what we do now is eventually added or don't know or are neutral. We debate this a lot. Using the four point scale is a forced choice. In other words you have to agree or disagree. You can't go in the middle. And if you put a middle point in there more people will use that middle point. And so the question is do you want to force them to [Inaudible] or do you want to give them an out. Different people feel different ways about that. You can make up your own mind. Here's another thing, this is from Andy Perko [Assumed spelling] and his colleagues when he was at the University of California Berkley -- this is another civic responsibility survey. It is about community engagement. You can see that the questions that he's asking are different than the ones that I gave you before. They've got a lot the same kinds of ideas, but you see a different approach, more items, and if you look at the response categories, Andy's got six of them in there and he went into the slightly [Inaudible]. It's still forced choice, but he's giving a six point scale. Some people like these larger scales because then you can detect smaller differences. Some people don't like them because they say what's the difference really between slightly agree, agree, and strongly agree. There might be a better bigger difference between strongly agree and agree. And so when they look at the intervals they say these really aren't appropriate intervals because they're not the same length. Again, there's debate in the literature about this. Different people come down on this in different ways. And I'm happy to discuss this further if you want to know more about it. Here's yet another one. This is an academic engagement scale that we use. This one's got also very high internal reliability. And this one is related to pieces of that affective behavior and cognitive engagement. This one originally was done by some researchers at the University of Wisconsin. We have modified it over the years and used it a million times. And so you can see the different kinds of measures that we have just on that academic engagement piece. Letting on and I apologize for just dominating here, but in terms of sampling what we find is that in Learn and Serve America most people use the census approach, which means they have everybody answer the questions. The reason why people tend to do that is because they feel that's a fair way, that if people are accepting then they all ought to have the same requirements. That's an interesting way to go. It has fewer problems in one way, but it also has predictable problems. For example, when people don't answer surveys it's usually because they have a more negative response or they don't care. There are other reasons that people don't answer the survey. So when you're going to interpret -- when you have a census and you have a low response rate, you have to talk about that. And you have to talk about how you would interpret that low response rate. Representative samples are okay as long as you actually represent the entire population that you tend to generalize. There are specific steps you need to follow when you draw a sample that is representative. This goes way beyond what I'm going to be in this seminar, but in general you may end up with error bands on this and that you want to be very careful because when you interpret your results you'll need to be very good about the representation of the

people who are responding here. Again, with us, we feel that is really important to get these comparison groups -- it's really hard to get good matched comparison groups, but it's so worth the time for what you can get. With human project protection again this is incredibly important. Again, subject protections are those kinds of protections that say you are protecting particularly children, but it's the children [Inaudible] in your study. With human subject protection, you have to guarantee certain aspects of your study. The most important of which is how you treat people and how you treat their responses to your study. There are specific protocols that you have to follow for informing everybody in your study about the purpose of your study and how the information is going to be used. You have to also tell them about and follow very specific steps for preserving confidentiality. Or if you're not going to preserve confidentiality you have to tell them that at the very beginning and they have to sign off on this. And basically they have to sign off on anything anyway and I'll tell you about that in a minute. You have very specific things you need to do about treatment of data including who sees it, how it gets stored, whether people who identify, who can see names can also see the responses, and so on. There's lots of different tests. Those of you that work at universities, most of the universities have their own institutional review boards that you need to go through. For those of you in school districts, sometimes you have your own research committees you need to go through. And anybody else, there are commercial IRPs that you can go to. They're kind of expensive, but they'll review your protocols and they will fix them or tell you to fix them and approve them. And basically this helps you to certify that you're doing what you need to do. And there should do be a lawsuit against you this helps to defend you against that lawsuit. You should be -- if you receive federal funds from the corporation you really should be securing IRP approval. If you are working with young people ages 18 years old and younger, you should be obtaining parent consent. And there are particular things that must go in parent consent forms and you need to follow that precisely. Sometimes the language is really weird and you hate it, but that's the way it goes. You really need to protect yourself. You also need to get participant info. There should be a sign off on everything that you do, both with adults and with children, saying that they agree to be part of your study and that the study is voluntary in nature, and they can opt out anytime, and they can skip any question, and so forth. That's part of protection of human subjects and so that needs to be followed. It will affect your results. If you lack the parent consent for example those forms don't tend to come back in with out a huge marketing campaign, so your response rate goes down. Almost half your response rate tends to go down a bit and even with ascent your response rate goes down. You need to know the flaw, you know, the problems in the outcome of doing this, but it's still something that you need to do. In terms of your analysis, one of the things that we found that just drives us crazy is that sometimes people use the wrong statistics. And yet they have a great design, they've collected wonderful data, they summarized it well, descriptively, but then when they go to look at group differences they're using the wrong statistics. And so you might want to consult with people who know what they're doing here and make sure you are using the right ones because otherwise you'll be torn apart. We are using, in the cluster, repeated MANOVAs, which are multiple analyses of variance, so it looks at your differences over time. We also spend a good deal of time determining the effect sizes. We either use Cohen's D or Hedge's G, those of you who are statisticians know what I'm talking about. Those of you who are not you probably ought to look into that. What effect size does is actually reduce down and tell you what portion of the variation in your outcome is attributable to your [Inaudible]. And so we think that's really important to know and so we determine our effect sizes. We also examine those moderators because we know that a lot of them make a difference. So in an overall study when we just run the statistics pre-post, service learning versus treatment, sometimes we find nothing -- no difference between groups. Then we start controlling for quality. We're controlling for teacher experience or looking at differences between different student groups depending upon their demographics or their achievement. And the differences start to pop out in a really big way. And it gives you a much bigger understanding of what's going on with your service-learning program. We

enter the moderators and mediators. We also use covariants like gender because we know that service-learning affects males and females differentially. Age is also a covariant. We also know that there are differences there. So those of you who are doing work in the field you need to pay attention those types of things that are well established in the literature as making a difference in outcomes. With qualitative analysis, make sure you're doing appropriate data coding and reduction techniques and appropriate summary techniques, so that you don't inadvertently mislead or highlight things that aren't generalizable. And again we are strong proponents of triangulating everything -- using multiple methods to measure the same thing. It's just good practice and it gives you much more confidence in the results that you're achieving. Every report really needs to have these elements, I mean, this is obvious, but it needs an executive priority in terms of methods, results, conclusions, and recommendations. Most evaluation reports are organized around the evaluation questions. It just makes it easier for the reader to negotiate what's going on. Again, clarity and ease of understanding and making sure that it is both accurate in terms of what it says, but also it's helpful for improvement. So that if you happen to find no impact you can help them in exploration and try to figure out why. And can give people some good advice on how to improve their programs. If you find high impact, you can share those results and say, you know, the programs that do more of this tend to get more of that. And frankly, that's how we found out about duration and intensity. Because across multiple studies we were finding the longer young people were in program and the more intense their experiences were the higher the outcome. Similarly, with link to curriculum we found that when a teacher can articulate the standard that students were learning and especially when students could actually articulate the content standards they were learning, the outcomes in epidemics were much, much stronger. And so, I mean, it sounds like it's obvious, but too many people just don't do it. So use of these tasks, these outcomes to help people improve just a critical thing for us to do in the field. And again, in drawing conclusions it seems obvious, but your data should be related to your conclusions. So many people just ignore their data or choose to ignore things they didn't like to see or can't explain something. So it's important to make sure that you cover all of your findings in a thorough way. Just very briefly and I want to make a picture for people trying to aggregate some of the results we're seeing across the fields so we can make a stronger case for ourselves. In the studies we've been doing -- we've done about 40 so far, we consistently find, of course, that quality matters. But we're also seeing some new areas top right now and we think the reason that we're seeing these things is first we think we've got some better measures and second we think that quality of practice has improved enough that people are starting to be more intentional about putting certain things in their designs. And we're seeing it in the results. Engagement and achievement are actually going up and it's very gratifying to see this because it helps us to make sure that our funding will continue. We're also seeing wonderfully promising things in 21st century skills. What this will do, we think, is help us make the college readiness argument that service-learning is a key component if not one of the most important to prepare students well for post secondary education. And so it's been a really gratifying to see some of these results. We've done a few other things. We're trying really hard to help make the dropout prevention argument for people. Our data -- we're just waiting for more data on that. So we don't have as much to say about that as we want to. We can say that the engagement, academic engagement pieces factor are promising here, but we're waiting to get out actual dropout data, our graduation rates, and our credit accrual, and some other factors that we're looking at in order to help make that argument. Again, this is all about improving and I sort of talked about this before, but make sure that you've got these pieces in there. Part of this is a question that comes up very often, is to what extent are you allowed to talk. The evaluators and program designers and program implementers talk to each other. There are times when it's just critical to have a strong collaborative relationship and that's at the beginning of an evaluation. When you're deciding your questions, when you're coming up with your design, when you're doing the execution, and so forth. You have to work really closely together to have a deep understanding of the issues that each

side has and to come up with the very best design you can given the program objectives and given the cost and given the other points of considerations that you need to make. While you're collecting the data there really shouldn't be a whole lot of communication other than making sure that everything's on track. You don't want, as evaluators, you don't want to influence outcomes because it'd be attributed to you. And so other than your methods being done separately, data collection being done well, and you human subject protections done well, you really want to keep your influence to a bare minimum. And then once you've got your results you again should work very closely to talk about improvements, talk about what the data means, and so forth. There's a lot of great models for feedback, but again it's the relationship that really matters here. It's important to have a close relationship with your -- between these two groups. A couple resources for you and it looks like I'm only going to have a few minutes left for questions, but here are my two favorite logic model resources. They're slightly different from each other in the way they present logic models, but both of them are fabulous in terms of step by step, here's how you do it, here's what good ones look like, and so forth. And then some other resources that are available for you. There are some guides that are out there, including ones that we've written. There's a toolkit. It's an overall approach to evaluation for service-learning. It's getting a little old now. I think it was written in 2006 or so, but the same basics apply. If you need information on how to do a really good job at collecting focus group data, classroom observations, or constructing surveys we have products on our website that you can just download for free, that give you some good hints on that. Also on our website we have some samples of survey subscales that you can -- old surveys that you can access and those are on the companion [Inaudible] sent. And research tools cart. So again you can go to RMC's website, you'll find it there. The research hub at the National Service Learning Clearinghouse is really good, especially for those of you in higher ed. It's just got Bringle's tools and a lot of great information around what to do. And that said, I love Bob Bringle's book, his book written with his colleagues. He's from the University of Indiana. And this is called *The Measure of Service Learning: Research Scales to Assess Student Experiences*. They're all higher ed, but he's just got, you know, a hundred -- more than that pages of samples of it that you could use that have high reliability associated with them. So I'm going to turn this over to Scott.

>> All right, I can't believe there's still four minutes before we're done and you got through that entire slide presentation. It's great. I was thinking people are going to have to step away from this and reflect some -- use service-learning terminology. But before they do I wanted to talk about a couple of the broad categories of questions that people came up with. One was, do you think you can say a little more about getting evidence of change in communities and community impact and making those improvements on the community needs that we're encouraging people to take a look at. That was one that there was a good handful of folks were curious about.

>> Sure. Measuring community impact is really difficult. There's a couple of ways that you can do this. Some of which are more rigorous than others. One of our favorite ways that's really difficult to do, but it's starting to yield some interesting results, is to look at civic health and disease from a neighborhood. A lot of states have civic health indexes or indexes or an index that they use. Sometimes there are community health ways that are already there and so if you can grab the pertinent aspects of existing surveys you are -- it's like gold when you're doing community impact. If you can't do that and it is hard to do, then there are a few fallback positions. One is to conduct surveys among a community partners and community members. We're trying to do that with our cluster. I must say we're having more of a struggle in that area than almost any other area because it takes time and people are not quite invested in the same. However, if you can get the people who receive the service or the partners who can see the impact on their own organizations and so forth to respond to surveys -- if you can get even more objective data. If you can look at the difference in water quality, for example, over time or if you can

look at the difference in the types of services that elderly receive. If you can measure body mass index on a childhood obesity initiative. Or there's lots of other ways to measure physical fitness, but you don't want to go for here-- just to let you know, don't got for height and weight because that's too embarrassing on the individual level. Anyways, there are lots of different ways to think about this in terms of who the community is that's being served and what it is that you need to measure. So it's probably great if we would spend more time as a community to develop these things and share.

>> And I appreciate the tip on the civic engagement because the NCS has a partnership, folks may know, with the National Conference on Citizenship to sort of jointly produce that annual snapshot of civic health index from community to community. And so the possibility of that that will be aligned with service interventions and civic outcomes is pretty strong. Thanks for that partnership we have there. Speaking of thanks and partnership, let me say again how much we appreciate Shelley taking the time to address what turned out to be an audience of upwards of 90 folks today. And appreciate the great questions that people tossed out. And Liberty and her team at the Clearinghouse for providing the platform and hosting all the documents as well as the recording. So that goes to Larry, Liberty, Shelley. Again thanks to all and all who participated and hope everybody enjoys the rest of your day and week. Thanks.

[Background sound effects]

>> Thanks everyone.

[Background sound effects]