

**Panelists:**

M. H. Clark

Mary Kay Falconer

*Please note: The following is a direct transcription and has not been edited.*

---

Mary Kay Faulkner: Okay, good afternoon. We are going to go ahead and get started. So, the sooner we start the sooner we're done, right. You know, the last session of the day, it has been a very busy, active day I'm sure for everybody. Okay, I'm Mary Kay Faulkner and this is M. H. Clark. We are the team presenting to you this afternoon on topic of validity in quasi-experimental designs to determine home visiting program effectiveness. So, if you thought you were in different session, that's the time to move. The learning objectives for the session; let me go through those quickly here, reviewing threats to validity and evaluations for home visiting programs and actually more than that it's going to be a good review of validity in research. So, I think that's the way we'll probably respond to it, become familiar with two statistical techniques, in particular the minimized selection bias in quasi-experimental designs and how these statistical adjustments balance the treatment and control groups in a quasi-experimental design to obtain the less biased treatment estimates for a home visiting program and again this is going to be good for you beyond home visiting, it will help you with the way these techniques were applied and how they might apply in some other program evaluations.

Then the last thing, is understanding variations of propensity score use in how they might be used to supplement RCTs. We actually would look to kind of discuss that a little bit, maybe some of you had some experience with using propensity score adjustment with RCTs. I don't know how many of you actually conduct research. We have one researcher there or all of you program evaluators. Okay, anyone not a program evaluator? Okay all right. Okay. Just a couple of things about M. H. Clark and I, we met actually at the American Evaluation Association Conference a few years ago. I believe it was when it was in Orlando, no before that. Yeah, before that, when M. H. was doing a presentation on how do you actually do propensity score adjustment and she was the person they brought in to actually show you step by step using SPSS output, how to do it. So, I was working with a program at the time that had just completed and another evaluation of a home visiting program in Florida wanted to do an analysis that involved propensity score adjustments so she assisted with that and served as a consultant in that particular project. So, since that time she has been considered an expert with AEA. I don't know how many of you to go American Evaluation Association Conferences? Have any of you been there?

Okay. This year it is Anaheim and I won't be going but, M. H. Clark will be there servicing masters, teacher, oriental master session on propensity score adjustments, so you might keep that in mind the gentleman who is headed out there. The other thing I would like to mention is validity of the topic that never goes away. It's always around when you are talking about design, evaluation design. I don't know how many of you

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

have seen this particular publication, new directions, and evaluation. It's something that again AEA Publishers. The topic is the dancing validity and outcome evaluation, theory and practice and the way I do this is again the continuing debate and discussion on validity and how to strengthen validity, different types of validity what designs or help one, which one is hard, how we can improve overall validity in our research. So, the debate continues. So, if anyone hasn't seen this yet you might want to look at it. All right so M. H. is going to start this off here and go over, sort of a review on validity, different types of validity threats to validity. So, let her take it away at this point.

M. H. Clark: Thanks, Mary Kay. If I start talking a little fast it's because I'm trying to be aware of the time. I realize shortly before I came down here that I have got 30 minutes to give a lecture that I usually spend half a semester on. So, I will be brief on some points. If you have additional questions please feel free to ask but, I am going to go through some of the point rather quickly for sake of time. There are considered to be four types of validities. So, validity for those of you, I'm assuming most of you guys are familiar with validity. It's this idea that research is doing what it says or the claims that we make from research are reasonable claims. That is, if we conduct a particular study it's limited by its local and specific constraints but, can we generalize beyond the findings of that particular study to the broader population or broader constructs. In 1963 Julian Stanley and Don Campbell came up with a monograph that covered two types of validity and threats to validity and in their original publication they only focused on internal validity and external validity. Since then in 1979 Cook and Campbell came out with four different types of validity and in the latest version of validity and threats to validity Shadish, Cook and Campbell in 2002 also maintained those four different types but, they rearranged some of the threats to validity.

Since I am a student of Shadish I'm going to follow Shadish, Cook and Campbell's version of validity. However, I have noticed that a lot of research areas particularly in education, I work in education, a lot of them are still holding on to Campbell and Stanley's original two types to validity and their original threats. So if you hear some of the things that I say and they, well that's not how I remember it, if you just talk to the original Campbell and Stanley version that makes feel why there are some discrepancies. In the four types statistical conclusion validity emerge from internal validity and construct validity emerge from external validity. Statistical conclusion validity is the first type and that addresses how the cause and effect co-vary, that is, it's particularly concerned with statistical relationships. If we conduct a statistic looking at the relationship between two or more variables those of interest is it reasonable to make conclusions from those particular statistics?

The second one is internal validity. Internal validity address whether or not the relationship between variables is specifically causal. So, a lot of times when we conduct a study we may be interested in just looking at the correlation between variables or factors but, we can't always establish that the variables are causal in nature, does "A" cause "B" or are "A" and "B" simply related. So, internal validity is concerned with whether or not we can determine specifically causal relationships. Construct validity is concerned with whether or not we are measuring what we intend to measure. You may

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

be familiar with construct validity and the context of measurement in, which are we specifically measuring what we say we're going to measure and the context of research. This takes on a more general definition. So, construct validity is interested in whether or not we're operationalizing the constructs appropriately. So, if we say that we are conducting an intervention on, that will reduce child abuse and neglect is whatever outcomes we use going to be a reasonable representation of that construct.

Likewise, if we say that we're going to conduct a cognitive behavioral therapy or the procedures that we are using for that particular type of therapy really representative of that particular type of therapy. So, are we manipulating or measuring what we claim that we are that we are interested in. External validity is interested in how well we can generalize from the specific operations to the broader constructs. So, if we take a sample from a population is it reasonable to then generalize to the population from our sample? If we use a very specific treatment manual is it reasonable to say that the effects that we find are going to be similar if we use a different treatment manual but, still the same general treatment. So, are we able to generalize from our particular study to broader areas of interest within each of these four types of validity there are several threats and according to Shadish, Cook and Campbell there are 42 specific threats to validity. Yeah, so I'm not, yeah, yeah, 42 its all like, you know, in earlier versions Campbell and Stanley they have much fewer and so of course my poor students get confused when they open their text book and they've got you know, , like 15 and I say, "No, you have to learn 42."

I'm not going to cover all 42 and I even had a hard time restricting myself to what's on here and I'm not going to go into a lot of detail even with what's on here. This only represents a sample from each of the four types. But I did point on the ones that I see most often in evaluation. So, within statistical conclusion validity and remember this is concerned with statistical relationships. We're worried about low power, low statistical power. So, power is the ability to detect an effect or relationship between your variables if one exists in the population. So, some things that may impact power are sample size if we have a really small sample size we may not find an effect that really exists simply because we don't have enough participants. So, that's one aspect of low power. Violating statistical assumptions, now you guys may have to go back to when you were studying research and evaluation and go back to perhaps nightmarish classes on statistics where you are taught, "Okay, if you're going to use this particular statistic you need to make sure your data meet particular requirements or assumptions and if you don't meet these assumptions then whatever statistic you come up with may not really have an impact, may not really be meaningful."

So, did you check to see whether or not your data meet those particular statistical assumptions? By the way I found a lot of people don't check. Fishing for significant effects, a lot of times we were having a hard time finding something that, is statistically significant we go fishing that is we start looking for a whole bunch of different relationships until we find something. The problem with this is that you're going to find something that is statistically significant by chance alone, even if that effect doesn't exist in population and the problem is by fishing you increase the likelihood that you're going to find an effect that doesn't really exist in a population. Unreliability of a measure, I

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

think it stands to reason, that if the measure that you're using to measure an outcome isn't reliable or valid, if it's not giving you a consistent result, any statistics that you're going to get from that measure is going to be something that you can rely on. Restriction of range, this occurs when there is a much broader range of outcomes in the population than what you collect from your sample.

And I ran into this several years ago with a student in her thesis where we realize that student drinking range from students drink it all to the point where they're drinking 40 drinks a week, they were alcoholics, even as college students. However in her particular sample she only had students who were low to moderate drinkers. This was not an accurate representation of her population. So, the problem is that when we have a restriction of range then we may not be able to find an effect that exists in the population simply because we don't have enough variability in that particular sample. Threats, with respect to internal validity, my personal favorite ambiguous to *[indiscernible]* [00:13:11] precedence a lot of times people want to make causal relationships when in fact they don't whether the cause proceeds the effect I once had a researcher who said, "Oh, my goodness, I cannot believe that parents yelling at kids causes them to smoke." More, likely the parents yelling at the kids about smoking because they knew the kids were smoking. So, the problem is that if you don't know whether the cause came after, you know whether the cause came first and then the effect; you can't really establish a causal relationship. It could simply be co-relational.

Attrition is when participants drop out of treatment conditions. Maturation is when there is a natural progression or a natural change despite the intervention and I think this is something that is of real concern with people who are evaluating children and child welfare programs because of course the children themselves are going to naturally progress despite your intervention. History is when something else occurs at the same time as the intervention that may also impact the outcome. Trusting is a type of practice effect. I don't know how often you guys will see that in this particular field but, I know in education this is something we see a lot, that is people learn to answer the test or change their answers on the test because they're learning the test by taking it multiple times. Probably the one that we're going to talk about more than anything else in this presentation is my personal favorite, this is Selection. Selection is non-random assignment. It means that if the participants are not randomly assigned then it's very likely that there are differences between our treatment and control group before the intervention. If we know that there are pre-existing differences before the intervention, well of course there are going to be differences after the intervention but, that's not necessarily because of the impact of the intervention and this is one of the things that we're most concerned with in this particular study.

Other things that we're also interested in with respect to construct validity one threat, is reactivity to the experimental situation and this is also something that I'm really concerned about with this particular field. This threat is when participants respond to an experiment or intervention or evaluation in a way that they think is more favorable or puts them in a better light. Basically they know that they're being experimented on or they know they are being evaluated. So, they're going to modify their answers or their

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

responses in a way that suits them. Experimenter biases, is another thing where we now have to worry about the experimenter interpreting responses in a certain way. So, as an experimenter if I know you're in the intervention group and you give me a response I may interpret your response as a way that supports the treatment because I know you are in the treatment program. So, for instance if I'm asking parents about their behavior towards their children.

And I know that they've been in treatment or they've undergone their part of an intervention then I'm likely to view their responses in a more positive way simply because I know they're in intervention group. Treatment diffusion, this occurs when people in the control group learn about the treatment and start using it themselves. So, if I put the Smith family in the control group but, they learn about the intervention from the Ortese family then it's likely the Smith family will start using it and even though they're labeled as being in the control group they're actually using the intervention, they've been exposed to it. Construct founding this is when I think I'm measuring just one construct. For instance, I may think I'm just measuring child abuse where I may actually be measuring both child abuse and neglect but, I'm not making a distinction between those two different constructs. External validity, main threats or interactions of the causal treatment with units, treatments, outcomes or settings and this means that I may not be able to generalize from particular people in my sample to another sub-population or population, or I may not be able to generalize from the specific treatment that I've used to other types of treatments, so for instance, in this case home visiting programs.

We may find that the results from our specific home visiting programs do not generalize to other home visiting programs. Outcomes, this is one of the concerns that we did have at this particular study and that is many of other evaluations of this kind of program use self-report. Well, if we use stage reports, records from state records we may find that we can't generalize the results from our study to another that use self-reports. Or, of course in settings, so for instance, if we conducted this as a field setting, that is, we went to people's homes to conduct an evaluation we may get different results than if we had conducted this in, say, a safe house. So what do we in particular need to become strong with respect to each of these types of validity when it comes to these types of evaluations? So, with respect to statistical conclusion validity the first thing, that we want to ask are what are the most appropriate statistics to, the more appropriate statistics when evaluating program outcomes? And I will always certainly saying the same thing. When you are trying to determine your statistics or any time you are setting up an evaluation or research the statistics depend on the design, the particular comparisons that you're making and of course the design, there is comparisons that you're making depending on the research question.

Everything begins on the research question for the purpose of the evaluation. That's what's going to determine your statistics. In terms of the relationship between the statistics and the design, these are a couple of things that do need to be taken into consideration in terms of using appropriate statistics. When the samples are related you need to make sure that you're using an appropriate statistic or within subject statistics. So, for instance, a lot of times people don't want to assign people to treatment in control

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

groups rather they would observe the behavior before an intervention have everyone exposed to the intervention and then observe them again after the intervention to see if the intervention was effective by comparing the pre-test and post-test scores. Perfectly reasonable design, but, make sure that if you're measuring the same people at pre-test and post-test that you're going to use a statistic that takes that into consideration. So, there is dependency between the pre-test and post-test scores.

So you will need to use a within subject statistic something like a related measures t-test. When observations are related you will need to use another type of statistics that will also take that into consideration. I've noticed that particularly since working in education this is an issue. However I can definitely see that in this field where if you are taking measures from not just one person in the family but, from several people in the family. So, for instance, if you're taking measures on several children within one family it's quite likely you're going to get some relationship between the answers of those siblings. So, I'm pretty sure that if you've got four children all in the same family then the responses that those four children have are probably going to be more similar to each other than children from another family. There are certain statistics that take that into consideration, the similarity between the families. Make sure that you use a statistic that does that. As I said before often when we violate statistical assumptions people don't realize this mainly because they don't check and I learned happily to get into this habit myself from my students when I realized my students were doing research papers where they were checking all their statistical assumptions and I thought, "Wow, they're actually listening to what I was saying when I was teaching them statistics."

Knowing that you're meeting the statistical assumptions ensures that the results that you're getting from those statistics are valid. Do we have enough data? Now, I'm still learning more about the particular samples in this field and the sample that Mary Kay brought to me several years ago, had a lot of participants in it but, a lot of the data that I see as a consultant often does not. So, very often I will work with students who are interested in specific populations. One of the big problems that we've run into is that we don't have enough participants and if we don't have enough participants then we're not going to find a significant effect even if such an effect exists in the population. We also may run into problems and this was an issue that we ran into is that there is a lot, sometimes there is a lot of missing data. So, if you have evaluations where you're looking at people over time and you may find when you try to observe that multiple times that they're not available or they're not going to respond at certain times or they may not give you particular types of, pieces of information.

Sometimes because they don't have the information, sometimes because they're not comfortable giving you that information. So, sometimes you may not have enough data simply because the observations aren't there, not just because you didn't recruit sufficiently. I noticed that in the, the early sessions this morning when we were talking about internal validity, particular randomized control trials and Mary Kay did warn me that she says this is very popular here and I said, "Okay, that's, that's perfectly understandable." People want to have efficacy trials that use randomized control trials. I understand that and I'm certainly not making an argument against them however I will

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

warn you that just because you have a randomized design does not mean you have ruled out all threats to internal validity and this is one of the big mistakes that a lot of people make. I was very pleased when, I think it was Dr. Shaw this morning said, “Yeah, randomized control trials are great but, sometimes they can also be misused.” And this one of the problems that a lot of people run into and that, is they think because they have used a randomized control trial they have solved all of their research problems.

You do rule out many threats to internal validity but, randomized control trials are not completely invaluable. They do rule out threats to selection so obviously people are not self-selecting into programs when they are randomly assigned and you do rule out interactions with selection. It’s likely to reduce the impact of other threats such as maturation in history. So, for instance, if I randomly assign my children to treatment in control groups, yes, that is children probably will mature and change their outcomes as a result of that maturation but, if the children in the treatment group are maturing the children in the control group are probably going to mature at the same rate. So, random assignment does help balance certain threats. However be aware that randomized control trials are still susceptible to attrition and I like an, attrition to self-selection in that if we consider that self-selection is participants intentionally selecting into treatments I think of attrition as people self-selecting out of treatments. So I maybe able to control people selecting into treatments by randomly assigning them to groups but, that doesn’t mean that they’re going to stay in those particular groups.

So, it is possible that non-randomized groups can be balanced not just with respect to attrition but, with respect to other things and this is a big concern that we may have is that we, reasonably assume that if people self-selecting the treatment then they’re going to be somehow different. Those in the treatment group are going to be different than those in the control and we should make that assumption. With random assignment we assume that if we randomly assign people to groups they’re going to be the same on all possible variables. They’re going to be completely balanced. And if random assignment does what it’s supposed to, then that’s a reasonable assumption, the problem is sometimes it doesn’t and I recently ran into this with a student who had very small sample size. She was looking at treatment for students with Asperger’s syndrome and she randomly assigned her participants and it wasn’t after she administered the intervention that she later realized that they were not the same on the pre-test. So, when she looked at performance scores at pre-test she saw that they were significantly different even though she had randomly assigned. There is no reason that those scores should be different but, for whatever reason they were.

So it is quite possible that when we have small sample sizes when there is a strong preference for treatment or when we have attrition then those randomized design cannot be balanced. Sometimes we can use designs without control groups. So, for instance, or I should say using designs without separate control groups and they can reduce certain threats to validity but, they’ll likely increase threats to internal validity. So, for instance, if we use a pre-test/post-test design where everybody is measured before the intervention, everybody receives the intervention and then we measure the outcome again after the intervention, it’s going to reduce certain threats like treatment diffusion, it’s going to

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

increase power but, it's going to increase testing, so everybody is going to learn the test. If history is a problem, if maturation is a problem without a separate control group I may never know whether or not those are impacting my effects. And finally something else that was talked about this morning that I really liked and we try to put this into our presentation a little more and that's model fidelity.

So, I found that I can come up with a beautiful design, I can have a wonderful cross-design with several observations over time and everybody shows up and everybody is supposed to be in the group that they're supposed to be in and for some reason these participants just do not get the message, they do not follow my design. Sometimes it's the administrators who don't follow the designs. Sometimes if I say that I want Marko to be in the control group but, then administrator says, "You know, Marko really needs the treatment, I'm going to put him in the treatment." That messes up my design now it certainly does give Markel that advantage of receiving the treatment that he needs. But as an evaluator that's going to give me some problems in terms of my treatment effects. With respect to construct validity we need to be concerned what, is the purpose of the program. So, again are we measuring the intended outcome a lot of times when programs are evaluated we don't focus on the outcome. What is the program supposed to be solving, what problem is it supposed to be solving and is that the problem that we're actually addressing and I've seen several evaluations that don't do that. What data are available to measure for this particular progress?

So, a lot of times we'll have data that are given to us and somebody says here evaluate this figure out of our program is good. So, we may not have control of the data collection. If that's the case do the groups include a reasonable counterfactual? The counterfactual is what would have happened in the absence of the treatment. Now clearly we can't have a true counterfactual, that is, we don't know that would have happened to the treatment group if they hadn't received the treatment because they received the treatment. So, instead what we try to do is construct something that's going to be the best representation of the counterfactual if what would have normally happened to this treatment group is an alternative treatment, not a treatment of interest then having a no treatment group may not be your best solution. But if it's likely that these people that were interested in providing services to normally would not have services then that would be a reasonable comparison group. Do the data sets include all relevant information? So, it may be that we're missing certain pieces of information that we don't have sufficient data.

In the case where we're using achievable data or someone else is controlling the data collection we may not have much control over this. Are the available measures valid and reliable? Again if someone else is collecting the data or even if we're collecting the data ourselves do we know that what we're measuring as our outcome is reasonably measured and likewise have we developed an intervention that addresses the need that we're trying to meet. Are we going to get valid responses? And again this is a big concern that I have with a lot of the programs that I have helped evaluate in the past is whether or not the participants will be honest. Especially when we're talking about something that's very sensitive such as child abuse and neglect. So, if I ask my, if I go visit these homes and I

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

ask my participants, “So, did you leave your child unattended for, you know, any period of time in the last month?” It’s quite possible that these participants are not going to give us an honest response. It’s also possible that they may not remember. So, if I ask how often, did you leave your child unattended in the last month they may not be able to remember that. So, a lot of times participants may not be honest with their responses or they may not be able to accurately recall behavior.

With respect to external validity we need to be concerned with whether or not we can get a reasonable representation of our participants and I know that particularly in our State, we are from Florida, we’ve got a pretty diverse population in the State of Florida. So, can we obtain data from the population of interest with respect to cultural orientation, racial diversity, language barriers and I know that Florida is not the only state that has to be concerned with these things but, having come from a different State recently I’ve noticed that Florida is much more diverse in a lot of these things. We have a very large Spanish speaking population. If our evaluators don’t speak Spanish we’re very likely going to miss a pretty big chunk of our population and I know that also in different regions of the State we have different proportions of racial diversity.

If we, if we’re only going to measure white families then we’ve got a serious problem particularly for people in, say, Orange Colony where only 50% of the population is white. Consider the stage of program or replication, so if you’re interested in making generalizations from one program evaluation to another consider where that program is in terms of its conception development. If you’re evaluating a program that’s been around for say ten years you’re probably going to get a different result than if you’re evaluating the program that’s only been around for one or two years. Also consider things like multiple program sites that are operating. If you’ve got a program that’s only operating in one particular location then you’re probably not going to generalize to other locations. However if you’ve got a program that’s operating in several parts of the State or several parts of the country then it maybe reasonable to generalize from that particular location to other locations.

Also consider the method of observation and I did mention this earlier. If your observation is self-report versus other report versus an observation then you may get different results and as I said this is a concern that we did have with this particular evaluation because we use a type of other report, which means that someone else’s reporting behavior on an individual, not the participant him or her herself. So, if we’re using State records are we going to get a different result than if we ask the participants to self-report on their own behavior and finally with respect to settings the results vary by region, this particular evaluation was looking at home visiting project program in Florida are we going to find the same results in different States.

Would we find the same result in different areas of Florida, so South Florida, Central Florida and the Pan Handle are culturally rather different. So, are we going to find consistent results across these regions the first thing we wanted to do is we wanted to focus on all four types of validity and we do want to emphasize that all four types of validity are important and one of the reasons that Mary Kay mentioned, the new

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

directions of program evaluation is that this series covered all types validity, not just in research but, also in evaluation and of course you'll see that there is a strong overlap between these two fields. Our concern even though, we think all four of them are important because we used a quasi-experiment design, we realize that we are most threatened by internal validity.

So, we're wondering whether or not the internal validity of our study is still going to be good or reasonable given the fact that we did not randomly assign participants to groups. Knowing that particular threat, without the statistical adjustment if we were to evaluate this as a non-randomized design without considering any statistical adjustments is that going to be a reasonable analysis? Probably not, as I said before and I don't think I have to convince this audience that any effects that we find from a non-randomized experiment are likely to be biased that is those effects may be due to pre-existing differences. If my participants are self-selecting into treatment programs it's very likely that those whose are selecting into a treatment program are probably more motivated to make improvements in their family situation than those people who don't. So, perhaps the motivation alone or the will to do better is going to be what's responsible for any positive outcomes we see that we may mistakenly attribute to the treatment of the program intervention.

Not considering statistical adjustments or not using statistical adjustments maybe appropriate if we don't have any measure of these potential biases. Now this is not to say that we should put a lot of stock in whatever results we find from this kind of design but, we may still be able to report something and gain some knowledge of a program that didn't use randomized designs but, I wouldn't like as I said I wouldn't put any stock in it. A better idea is to use some sort of statistical adjustment to get a better estimate of those treatment effects and two of these types of adjustments are co-varied adjustments and propensity score adjustments. With co-varied adjustments known biases are measured and included in statistical models. So, usually what happens is that we conduct we collect information on other characteristics besides the outcome measure and the intervention. Things like age of the parents, risk of child abuse and neglect, education, these sorts of things that we know will contribute to the particular outcome and possibly that will also contribute to whether or not they will participate in the program and we can take these particular covariates, put them in a statistical model and get a more accurate representation or more accurate relationship between the intervention and the outcome.

Certain types of statistics that will allow us to do this, we can use an ANCOVA in, which the covariates are the covariates we can use multiple regressions and we can use matched analysis and, which were matching on those covariates. With propensity score analysis or propensity score adjustments we're going to use the same basic statistics such as ANCOVA's multiple regressions and matched analysis but, we're going to take multiple covariates and aggregate them into a single covariate. So, this is very similar with the covariant adjustments but, now we're going to basically pool all of these covariates into a single variable. The propensity score is the predicted probability that a unit or a participant. And I say unit, because it may not necessarily be a child, it could be a family unit, will be assigned to a particular treatment group and it could be that they self-select

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

to. So, what is the likelihood given a set of known covariates that this person or family will select into a treatment. The particular statistical analysis is going to be very similar to covariant adjustment so I can still use propensity scores in an ANCOVA, in a multiple regression and I can match on propensity scores, which is very popular. But one of the other advantage is that uses fewer degrees of freedom.

If my goal is to equate or balance my treatment and control group then I can use the same covariates in the covariant adjustment and the propensity score adjustment although very often the goals are slightly different. So, for instance, with the propensity score adjustment it's usually intended specifically to balance the group where a lot of times with covariant adjustments it's to get a more accurate relationship between the treatment and the control group. When we are selecting covariates for covariant adjustment first of all the known biases need to be included in the model and this is true for both covariant adjustments and propensity scores. So, we need to know what contributes to the bias. So, if we know that people are self-selecting into treatment conditions and motivation is a big factor we need to try to get a measure of motivation and include that in our model. Make sure that you measure these covariates. Are you going to be able to measure them? This is one of the potential problems that we may run into it is that we may come up with a great idea of what may contribute to self-selection into treatment programs but, that doesn't necessarily mean that we're going to get that information. Will participants provide this information and are they going to accurately give us this information?

Also make sure that non-ignorable covariates are accounted for. Non-ignorable covariates mean that we need to pay attention to these covariates. These covariates are things that do make a difference in terms of whether or not people select into a treatment program or not. So, if we don't include certain variables that are related to selection then we may not be able to get reasonable treatment effects. What are some of the basic steps with covariant adjustments? Well, first of all the hardest part is obtaining the measures of the covariates or we may have problems in terms of getting samples, we may run into problems of measurement attrition, that is, people don't actually give us the data that we need or people may not report honestly. Once we have those particular sets of covariates all we need to do is add these into statistical models. The statistical models I mentioned include ANCOVA's multiple regressions. In some cases we may use matched T-test, blocking, factorial ANOVAs, again talking about the statistics, this is an entirely different lecture but, these are some examples of using covariant adjustments.

Because propensity score adjustments or propensity scores would be a type of covariant adjustment if your goal is to balance the treatment in control groups you could simply use the same covariates. So, the analyses are going to be the same, I'm still going to use things like ANCOVA or matching or blocking. The difference is that instead of using a whole bunch of individual covariates I'm going to use the propensity score. So to compute the propensity scores I'm going to select covariates that are related to both the outcome and selection and I have to admit that there is a bit of a debate on this. Accordingly to Donald Rubin you don't need to be worried about covariates that are related to the outcome. You just need to be worried about covariates that are related to selection because that's what causing imbalance. However if the covariates are not

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

related to the outcome it's not really going to make much of a difference. The covariates do not have to be significantly different at  $p < 0.5$ . You may set a more lenient criteria and that's fine. It will still add to that variation estimate the probably that each units or each participant will be in the treatment group from all these covariates using logistic regression.

Basically what you're looking for is a probability something between zero and one that will say the likelihood that this person will choose the treatment is this. The numbers that are closer to one suggest that participants are more likely to select the treatment, the closer the numbers are to zero suggest that the less likely the participants will select that treatment. Test the propensity scores for balance. Basically balance means that are the propensity scores and of course the covariates evenly distributed between the treatment and control groups. Okay, I went over longer than I expected but, so now that you understand the basics of the validity and propensity scores and the covariant adjustments this is how we use them in this particular study. So, I'm going to give this back over to Mary Kay to tell you how we use these concepts in our particular evaluation.

Mary Kay Faulkner: Thank you, MH. All right, the illustration here is Healthy Families Florida home visiting program. Most of you, several of you may be familiar with Healthy Families American model. I don't know does anyone know it. It's a long-term home visiting program, voluntary, it serves families that sets at high risk of child abuse and neglect, the outcome that is usually of first interest is reduction in child abuse and neglect, in crimes of child abuse and neglect, there certainly are other outcomes that are important for Healthy Families America model programs but, this is the one that we are focused on for this particular exercise. The evaluation that was used was a quasi-experimental design. I was not the evaluator. I was brought in late in the process and was to interpret the results and help with basically reporting results to the community and also do ongoing evaluation on the program and in this case again we did more analysis. This was moving into the, let's look at this again. The evaluation, original evaluation began in 1998-1999. There were 24 projects already in existence and then the evaluation ended in 2003-2004, 38 projects. So, a couple of things you need to realize and again these are comments that have been mentioned earlier.

This was an example of our program and it was highly successful, very strong infrastructure, very well managed, moved quickly, grew quickly and basically grew with the evaluation team that worked on the original evaluation in the way of developing a database that was quite sophisticated, collect participant data, which was then used in the evaluation, developing a good relationship with the State, maltreatment records that were available. So, again change occurred quickly with this program in the way of growth. The three types of validity that I primarily concentrated on now M H, has explained to you again the four types. I guess conclusion validity is statistical conclusion validity, I sort of included and internal and construct and others as we've looked at it but, primarily considering internal validity as M. H. mentioned the key problem here is that we do not do the randomized assignment so we were certainly concerned that there might be selection bias in the results but, a couple of things that were important to remember the participants in this evaluation were all eligible for services, which is important to know

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

and then all participants in the evaluation volunteered to be in the program but, were unable to be served due to what we call situation capacity and that is just not capacity to serve a program, it's when you situation-ally do not have the staff that has the right language to work with a family or the caseloads that are required for your family home visitors are exceeded and you can't add another family at that time.

So, a situational capacity becomes a very important challenge and so the control group that we formed from this original evaluation came from those who could not be served due to situational capacity. So, in other words they had volunteered to be in a program and they were eligible for the program obviously. External validity, as M. H. said in Florida again dealing with very diverse population throughout the State, cultural language, ethnicity being extremely important, the type of community they were in and certainly the racial composition. Construct validity, when the program was undergoing this evaluation, they did not have a self-report tool that was used now. In some of the home visiting program evaluations that you may be familiar with there are some other tools that are available, conflict tactic scale being one of them. There was nothing in place in Florida that was self-report and by contract the program had to rely on State maltreatment records. So, that was the choice for the outcome measurement with State, not treatment records.

Talking a little bit about design here and I'm sorry I'm not able to think help you deal with this and I don't want to your eyes to glaze over and all those but, the design was fairly sophisticated. I would consider this again to be a retrospective evaluation, not prospective and generally the prospective that's what it is recommended and preferred but, the data were there, evaluation was already completed and we went back in and used the data as that had been formed in the different multiple groups. And by groups I list them there, we had a no service group and again this refers back to those who are eligible for the program volunteer to participate but, could not be served due to the situational capacity, program completers was another group and then a group that was formed based on the their high fidelity experience in the program and Dugan, who was an evaluator of a couple of other home visiting program actually developed to a construct, they referred to a high dose or high level of participation being those that had 75% or more of their completed visits, home visits or what we called Level X or out of contact being they limited to a short period of time, three months. So, in other words there were criteria used to set that particular group.

While service does age these were participants that had been in a very short period of time, less than three months of services. In addition we had some sub-groups that we referred to timeframes, children that were up to 12 months of age and then there was another sub-group that was children up to 24 months of age. Now in the original evaluation we had a lot more than that. We had a 3-month, a 6-month, you know, I think a 36-month. So, there were more groups involved in the original evaluation. For the exercise that we went through was the secondary analysis, the additional analysis we limited to just the 12-month and then 24-month group. So, this gives you a table that sort of lays out what I just mentioned as far as the group comparisons and then also the 2-, the 12-month and the 24-month age groups and gives you the numbers that participate in

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

each of those. So, as M. H. was mentioning before it really helps to have high numbers of participants and this evaluation definitely met that criterion we had high numbers and so that again allow us to do a little more statistically than you might be able to do ordinarily. Okay. So, steps in this impact analysis. No statistical adjustments, which was one of the options that you included in an earlier slide.

Just looking at the percentage of children in each group that were maltreated we computed effect sizes and then basically this binary logistic regression was used where the dependant variable is the occurrence of maltreatment and then the comparison group membership in the model, Model one. It's the only variable in the model. Okay. And then what you are computing from there, it's an odds ratio. So, I don't know how many of you are familiar with binary logistic regression but, that is what is used and this has been used in other evaluations as well. Have you used binary logistic regression before? Okay. So this is the table that includes the effect sizes, with the percentages there for the different group comparisons, no service high fidelity, no service completers, low dosage completers group, low dosage high fidelity. I can provide the final that was used to compute the effect size of your interested and now I can send you the Excel Spreadsheet but, as you can see and again I don't know you're familiar already with effect sizes but, there are actually some pretty high effect sizes and I was surprised to get anything over 0.5, to tell you the truth so, with the low dosage service group and the completed group 0.84, 0.74, 0.65a, so, you know, we were seeing some sizeable effect sizes but, also some that were very low.

All right, covariant adjustments, what we did with this part of the analysis is develop different models and building on what we did initially with the no statistical adjustments. So, Model two actually included in each of that score and this is a risk assessment that is done for eligibility and admission into the program. If you score 13 or higher you're in the program based on the criteria that are laid out that is part of the scoring for this. It's really a unique tool. I don't it's done the same way in any other home visiting program that we're aware of fairly complete and comprehensive assessment. So, that was the only thing that was added to the model in Model two. Then in Model three we subsequently added more covariates. Then in Model four more covariates and some of the variables that are mentioned there, the covariates that you're seeing should be familiar to you if you're in the literature and home visiting and child maltreatment. Certainly those are important characteristics, those economic factors that should be included in any model and looking at these causal relationships.

So, we had a series of models and one of the reasons we did that was a problem that came up and was mentioned by M. H. already and that was we knew we had some missing data and we did not want to select one model for this particular exercise with just one set of variables because we would have too much missing data in a model that included a high number of variables but, on the other hand we were interested in what happened with some of those models that had more covariates. So, that was the reason why we structured it that way having different models that we were testing as we went along. So in the tables that we presented and by the way there is publication for this Children and Youth Services review, which lays out the documentation much better than we are doing

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

in this PowerPoint in this session of the day but, this is an example of the table that is displayed there showing for this particular comparison, which is the no service group and the completers group, the odds ratios for child abuse and neglect by each model. Okay, so when you don't have any statistical adjustments what you're getting is Model one.

So, we're actually replicating what was in the original set of results from the evaluation in the Model one findings and the way we interpreted that and M. H. correct me again if I'm wrong here before a point for 69, four times more likely to have maltreatment in the no service group. That's that way that would have been interpreted. So, it worked well for us to have it structured that way because in explaining the results out in the communities more likely to maltreatment in a no service group. So, the way that the coding went the way we structured this I think helped to explain the results. For each model it looks like the results definitely were consistent there as far as being statistically significant. Okay, this is just another example I'm putting up on a screen. No service group and in this case it was the high fidelity group that I mentioned we used a set of criteria to from that group and in this case the level of significance varied some over the different models and I also included there and again when you look at the fine print in the PowerPoint you can see how the N changes with each model and that's an important part of the challenges we had with the data we were using in this particular design because what was happening then by the time you got to Model four that is not what you see up there. What was happening to the N as you went along to a more sophisticated model with more covariates what was occurring. The N was dropping, yeah.

Okay, covariant selection I just mentioned there for the propensity score adjustments. We used I think the same factors are very similar to what we had with the actual covariant analysis. So, in fact I couldn't remember if we varied them at all. There might have been maybe one difference in the variables we used with propensity score adjustment computing the propensity scores that was regressing, the factors onto a dichotomous variable indicating treatment group membership. So, in this case again developing these variables that indicated what group they were in and then the predicted probabilities that M. H. had mentioned were products of this process and there was a transformation using a lot linear function and these values adjusted the propensity score. So, again it's a multi step process, it's something that is not easy to try to complete on your own without expertise in this field and strongly suggest that you get with an expert to work through it. The balance on the propensity scores M. H. had mentioned that she had this go through, had me go through all of the tests for each of the group comparisons using SPSS output. So, again a lot of guidance required in working through that. Among eight paired comparisons format of the criteria, which MH, I don't know how that compares with some other work that you've done, but, you know, I guess we were satisfied with that and thought that was good enough.

M. H. Clark: Yeah, that's pretty difficult especially when you have a lot of biases, there's a lot of imbalance between the two groups and you find that your covariates are very strong predictors of selection then you're going to have big differences between the propensity scores. So, unfortunately it's not unusual if they don't meet all of them.

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

Mary Kay Faulkner: So, the next step was the matching process and matching the participants on the propensity scores using the caliper matching procedure and there was no replacement in this process, there was actually a program written in C++, I guess it was the computer code that was used to write those and then in this case M. H. was helping us with determining the distance between the matchers, the caliper, I guess, actual value that had to be used in order to ensure the 95% reduction in selection bias. So, if you hadn't already gotten an impression here again having somebody who knows how to do this and works with you and trying to get it done. Okay, the last step was the McNamara test used to estimate the adjusted treatment effect for dichotomous outcome with match pairs and I'm sure M. H. could explain all a bit more to you about that, works like a chi-square for depending groups and then I'll show you here the results that were generated again appear in that article that I talked about. So, this was fun, you know, working through all of those but, it was an important exercise. And then we did some comparing of the results from the different techniques and just these are highlights from the things that we included in the article. There was certainly more consistency and the results across the techniques with them without adjustments then we expected, in other words, some differences but, more consistency I think than what we expected. And again I highlight some of the differences here, you know, statistically significant differences between the low-dosage service group and the high fidelity group for children and the 24-month subgroup in the statistically adjusted effects, both covariant and propensity score adjustments. So, there was a change when we used the statistical techniques.

Other group differences, statistically significant results across with and without statistically adjusted techniques, let's see what I'm grasping over that one. I guess just the level of statistical significance was varying a little bit within each technique and across the techniques, group differences increase but, not always significantly in most comparisons after accounting for covariates, variations in the results across the subgroups based on age of the children were not particularly that worrying, it didn't really matter that much if we were working with the 12-month group or the up to 24-month group and then, which ones had the lower treatment effects, propensity score adjustments had lower treatment effects when comparing no service and high fidelity and then up to 12 months and those service to completers and that up to 24 months. Traditional covariant adjustment had a lower treatment effect when comparing low service to high fidelity. So, again there were some differences but, not as much as we originally expected or I originally expected, I guess. The strengths in this particular exercise again are all participants being eligible for the services and volunteering to be in the program.

The covariates we had available to us were good, theoretical and empirical contributors to the outcome, particularly the age fact, the score. Using State records reduce missing data for the outcome manager and by that what we mean is that once we had that child in the database and that information there wasn't a problem with that child leaving the study. We could follow that child and whether or not there was maltreatment occurring with that child through the necessary time period because we have the State records. Now it's possible the family might have moved out of the State and then we might not have had it. But for the most part we feel that we had a fairly complete coverage of the outcome because of that, because of using State records. Okay, probably challenges and

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

limitations, limited number of covariates and again this non-ignorable variables or whatever, I mean that was big on the list and even though we had quite a few covariates that we could use for the traditional covariant analysis as well as for the propensity score adjustment we needed more.

So, those of you who are thinking about using some of these techniques intake or when you are bringing participants into a program make sure you have a really strong set of factors, variables that you can use for these types of statistical adjustments. We wish that we had had more with complete set of age fact responses or items for each, item of age fact, such as smoking during pregnancy was a big one. Another one was substance abuse. We had individual items on the age fact for a large percent of those that were in the study but, we missed it for a lot about there is because we didn't have the complete data on the assessment tool. It is just a total score. So, again what do you with that data? What you collect initially, how you maintain it or whether or not you have it later on just got to keep that in mind. The details are really important and can be very, very helpful in doing these types of statistical tests. I already mentioned about treatment attrition occurred only when the participants moved out of State on retention though we still had a high number of cases actually involved. Even though there was some missing data in some of the group comparisons we were able to keep a high percent in some of those comparisons, 95%.

So, we know it's pretty good and when you look at a randomized control trial where the problem like flies out there you know this was respectable. 95% of the participants that started stayed in and had complete sets of data so we are happy about that. The maltreatment measure we used and I don't know how many of you are really interested in this was more inclusive than as usually used. A lot of times they'll stay with what's called verified reports of abuse and neglect. In our State we were required actually by contract to use some indicators as well as verified and this was very important because it upped the number of children that we had in the dependent variable that had experienced maltreatment in many ways too that some indicators sort of serves as a precursor to more serious maltreatment. So, the measure of maltreatment was in other words broader, more comprehensive and worked better, I think, statistically with what we were doing. But a lot of the evaluations you look at home visiting they just used verified only.

Challenges and limitations, let's see, claims that the reports do not include all of the occurrences of child abuse and neglect with hospital records done including all of the occurrences of child abuse and neglect either and self-report, I mean, what do you think about self-report? Yeah, I do, no I don't, I mean, that's not and surveillance bias. Well, yeah, I mean there might have been some home visitors that reported maltreatment but, in this case I think it was good actually that that happened and probably makes it the results that we had stronger favoring the program if that was occurring the surveillance bias was occurring. State participants, from several projects throughout the State no participants were excluded due to language literacy skills or cultural barriers. When you look closely at some of the other evaluations that had been out there because there was a language barrier or a literacy issue, they could not complete the participants and the evaluation could not complete the questionnaires that were included for measuring a number of

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

different outcomes. So, that was a real disadvantage. In this case with us only being interested in child abuse and neglect we did not have any other tools that required literacy levels that are certain level or language, English proficiency. So, that was an advantage for us.

Community and project level factors could not be included, I mean, that was another unfortunate thing or limitation in this evaluation. And just home visiting in general, how challenging it can be to evaluate a long-term intervention very, very challenging and so we list a number of recommendations that you might consider in trying to do your own, quasi-experimental design using home visiting data and that it's out of time.

M. H. Clark: Another 15 minutes?

Mary Kay Faulkner: Yeah, I want to make sure she has an opportunity to get to regression discontinuity designs but, at this point is there anything in particular you want to ask about Healthy Families and the way of the structure this particular design and yes.

Audience: I want to ask a question about, could you tell me a little bit more about what you mean when said that use a little bit more inclusive maltreatment measure what does that mean?

Mary Kay Faulkner: Well, the difference between the evaluations that have been using only the verified reports as opposed to some indicators and verified.

Audience: What like a legal termination of sorts?

Mary Kay Faulkner: Preponderance of evidence is I believe that verified credible evidence is the sum indicator. Preponderance means an overwhelming evidence that maltreatment occurred whereas credible is... it is there but, it's considered some indicators.

Audience: Between substantiated and indicated or something like that?

Mary Kay Faulkner: Well, now that the terminology has changed even more than it used to be before this evaluation...

Audience: Yeah, it's different in every state too.

Mary Kay Faulkner: Now they're using verified and the others are not substantiated. So, the terminology has changed a little bit, verified I think is still consistent are you seeing states using what they call substantiated is that where it is...

Audience: Different yeah there are different terms they use and we're submitting it to N-Cans *[phonetic]* [01:13:30]. They fit it into the N-cans terminology but, that's not necessarily what the state is they have to sometimes reinterpret what they call it and bring into the N-Cans.

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

Mary Kay Faulkner: Well got to love that logic when the knocked out the sum indicators to send the N-cans only the verified reduction in maltreatment in the State of Florida might have been what people picked up. Whereas we monitor very closely the sum, what we still call sum indicators or non-substantiated and the verified we do that on an ongoing basis we look at it quarterly what's going on by projects and I look at list of maltreated children and see what the patterns are, what the, what's going on with those because we see the shift to verify it in some Counties or to not substantiate it not combining both when you look at county and you see only verified in another county only not substantiated then you know just the way they are doing the investigation and the way they are coming up with the finding and it is related to that particular county and maybe the training that has occurred or the supervision or whatever but, we still include both and we account for both. Okay, I think I may go and just briefly talk about regression discontinuity designs because that's another alternative that you might be interested in and then we'll wrap it up if need other questions.

M. H. Clark: The segue to this one of the reasons that we wanted to talk a little bit about regression discontinuity designs is this concern for doing randomized control trials versus doing quasi-experimentation. So, there is this tendency again if it's randomized control trial its real signs. If its quasi-experiment, you know, now you're just playing with numbers. Well I certainly don't think that that's true. I think that you can come up with some really good research designs with something other than a randomized control trial. One of the alternatives to randomized control trials in quasi-experiments is regression discontinuity design. And I'm starting to see these in evaluation much more. Really starting to see an explosion of these probably within the last three years in evaluation, I saw several of these at the American Evaluation Association Convention. And the regression discontinuity design is a very interesting way of using pre-existing conditions to control for non-randomized effects but, still allow people who need the treatment to get it. So, the problem with random assignment is that it's very likely that the people who need the treatment the most are not going to be the ones who get it whereas with regression discontinuity you're able to do that.

Regression discontinuity designs assigns participants to treatment groups based on a cut-off score. A cut-off score on a particular variable. So, the assignment variable that you use needs to be continuous. It doesn't necessary have to be related to the outcome but, it certainly can. So, for instance, income may not necessarily be related to child abuse and neglect but, you can still use income as a way of assigning people to whether they are in the treatment group with the control group. Essentially once I determine what a person's income is then I would determine a cut-off on that income. So, perhaps the mean income and say anybody who is above the mean income is going to be in the control group and anyone who is below the mean income is going to be assigned to the treatment group. So, the poor people are going to be in the treatment group. This can also help us give treatment to the more needy participants and of course if the assignment variable is related to the outcome or for the need for treatment then this is reasonable, a lot of times income is used as a, assignment variable. The nice thing about this is the evaluators have control over assignment. Like with randomized control trials we as researchers or

## Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness

evaluators have control over the assignment as opposed to self-selection where it's the participants who determine that. When the participants determine what groups they're in we don't know the assignment mechanism. We don't know why they're in that group.

But if we control it, if we say use income and say you're in this group because you were below the mean income, I know what assignment, what the assignment mechanism is and I can statistically control for it. So, I would essentially use that assignment variable as a single covariate. I know that there is bias in the treatment in control groups and that particular variable and so I can account for that. How would regression discontinuity designs work for home visiting programs and when Mary Kay first asked me about using regression discontinuity designs with home visiting programs I thought that this would be perfect, this would be a, this is a really nice design for home visiting programs, especially if we know that we can, if we know that randomized control trials can be used in home visiting projects but, it can be very difficult and regression discontinuity design is going to be very well suited for Healthy Florida Families. Now this is of course assuming that we can get the participants to stay in the assigned treatment groups. This is not to say, this doesn't rule out other things like treatment diffusion and other things like that. One of the things that we could use was this age fact risk score. So, the risks, the age fact score and correct me if I'm wrong on this Mary Kay is basically a way of assessing risks of child abuse and neglect. If we know what the risk is for child abuse and neglect then we can use that as the assignment variable, it a continuous score, so we can set the mean as the cut off and of course those people who are more, who are greater risk for child abuse and neglect can be assigned to the treatment group and those people at lower risk can be assigned the control group.

So, once we make that assignment then we could still account for that potential risk when we're estimating the treatment facts. So, we still have treatment in control. We do know that there was a bias but, we're going to be able to account for this bias. Unmeasured biases are assumed to co-vary with that risk score and therefore because if they do correlate with that risk score that's okay because we are accounting for that bias, we're accounting for the risk score then essentially we are also accounting for those biases that are correlated with that. Of course we still haven't been able to rule out all threats to validity with regression discontinuity design. Please be aware that no matter what kind of design we come up with it's not going to be perfect. We're always going to have problems no matter what kind of design we use or statistically use. Attrition can still be a problem with this but; hopefully we can reduce attrition if we are using people who needed the treatment more. Treatment diffusion of course can still be a problem but, hopefully we've diffused treatment diffusion by putting the people who need the treatment more in the treatment. But of course if these occur then we're still going to have, this is still going to adversely impact the validity of the results. Okay, I knew that we only have a few minutes left. So, are there any questions either about the evaluation or validity tests?

Audience: On slide twenty-three there was the final bullet it talked about that among the mean comparisons, they formed all three criteria. Explain that to me.

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

M. H. Clark: We didn't go into a lot of detail mainly because I would get completely carried away if Mary Kay let me. So, there are several, there is actually several more ways of determining balance. In 2001, Donald Ruben came up with an idea that if you test the balance of the propensity scores then this should also assess the balance of the covariates. So, remember the goal of randomized assignment is to balance the covariates and your treatment in control groups. We assume of course if random assignment works then all of the variables, the distributions of all possible variables observed and unobserved will be similar with the treatment in control groups. If the propensity scores are balanced then the covariates should also be balanced. So, these are the three ways that you can check to see if the propensity scores are balanced. The first one is to see if there is a significant difference or a large difference between the propensity scores in the treatment group and control group. So in this case we're just looking at the propensity scores. So, I would estimate the predicted probability for all of the participants and I would look at the mean of the propensity score in the treatment group and the mean of the propensity score in the control group. Now of course if my covariates are predicting selection and condition it makes sense that those people in the treatment group are going to have a higher propensity for being in the treatment group than those people in the control group.

This is fine that the treatment group has higher propensity score than the control group but, we don't want it to be a big difference because we want those propensity scores to overlap well. Unfortunately this is the one that's most often violated. So, we want to have a small difference in propensity score distributions and treatment control groups but, unfortunately it often is the case that we're going to have people in the treatment group have a much greater propensity for being in the treatment group. Group variances of the propensity scores are homogeneous and that means that we want to have distributions that are going to be, that are going to have the same sort of variability. So, if I have a nice normal distribution in my treatment group of propensity scores I want to have the same sort of variation in the control group. And the last one this is the trickiest one. The group variances of the residual errors after each covariant is regressed on the propensity scores are homogeneous and what this means is that what we want to do is we want to find the error between a single covariant and the propensity score. So, we're going to try to predict the covariant from the propensity score.

How much variance are we accounting for? So, we're looking at using the residual, how much error do we have. So, how much are we not accounting for in the propensity score but, how much are we not accounting for in that variable in the propensity score, so how much error do we have. And of course we expected that we're going to get a certain amount of error, we're not going to get perfect prediction between the propensity score and that single covariant but, we want to make sure that the error is evenly distributed in the treatment in control group. So, those are the three. There are other ways to test for balance. Usually people will simply look at after they do whatever adjustment. For instance, if they match on propensity scores or they block on propensity scores or whatever they do to try to make those balances then they look at the covariates again, not just looking at propensity scores but, they look at the differences between the covariates

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

themselves instead of just the propensity scores and that's, that's what they used to do and it's actually still a common way of testing for balance.

Audience: *[Indiscernible]* [01:25:29]?

Mary Kay Faulkner: Yes that's right the groups who are comparing and then the up to 12 months up to 24 months yeah.

M. H. Clark: So, unfortunately and, you know, Mary Kay had mentioned this that, you know, as well we didn't, we didn't talk as much about statistical conclusion validity and part of the reason for that is that well we have a lot of things that we can pat ourselves on the back for using McNamara tests instead of, you know, other types of tests, really considering that the matched samples, but, this was one of the places where our statistical conclusion validity was not as good, we, our validity was not as good here because of those eight pair of comparisons only four of them that offer a criteria but, the other one still met much of that criteria and the one that's most likely to be violated is that first one. But that's not unusual. So, further research that I've done suggests that violating that first one is not completely detrimental.

Mary Kay Faulkner: Yeah, I mean the important thing is to do it in a note that there may be an issue with it, whether or not they're totally invalidates the results, you know, guess that sort of subjective judgment but, you know, again the other thing we were interested in getting from you is whether or not you think these techniques are usable to do with RCDs and when you completed in RCD going back in if you have some concerns that your groups were not equivalent. Maybe you do some of these analyses to confirm, you know, to replicate your findings, you know, that's an advisable thing and I think we're going to see more of that down the road particularly with small N's although I guess there is one lady, she has a small hand, can't really do some statistical techniques as you require certain degrees of freedom.

M. H. Clark: Yeah, I mean, yeah. Yeah, I mean there are things that you can do. But I think probably the biggest thing is, I mean, was, you know, looking at a paper recently, it says, well, what happens if you find that, you know, you randomly assign people to groups and then you find that the groups are not balanced, what do you do? Well, ideally you randomize and keep doing it until you have balance. Most people don't do that. They don't have time for that. They don't stop and think, they think I used random assignment, I should have balance. Why am I going to go out and check for that now? And that's, that's reasonable because we expect random assignment to do its job. That's why we did it. Unfortunately sometimes, it doesn't and we need people to check on when we do have small sample sizes. This is one time in, which we may really need to be vigilant in checking to see if there is balance. But if we've already started the intervention after the random assignment, so we've checked and we found, "Okay, it's imbalanced. What can we do?" We can still do something about it. We can still make some adjustments and these techniques can still be used, propensity score adjustments still can be used when random assignment doesn't do its job such as whether it's even,

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

you know, either by some strange weird fluke that the groups weren't balanced, or because the groups were balanced at one point but, due to attrition they lost that balance.

Mary Kay Faulkner: So, American Evaluation Association Conference, close to hear these debates and they get more information on how to do these things and whether or not it matters if you meet all the criteria in your tests. So, I would encourage you again those of you who are interested in certain thing. Consider AEA as a source of information. Is there anything else, or anything you've noticed about what we've shared with you, you know, sometimes it concerns me greatly, the amount of intrusion that evaluation causes on a program that's being evaluated, how invasive you are with all the administrative tools and things you have to do with the participants and trying to remain as distant as detached as possible.

Audience: Well I'm thinking really on a more practical level after you are all done these are some heavy duty very detailed statistical procedures and approaches when designing but, when you get some results what you know, what can you do with this level of very, very detailed information about how it was done what do you say to the child welfare director or the governor, or to the legislators in a state about what public should we need based on your results because you can't I mean you may be evaluators here but, you can't say that to you know, the legislators, what kind of approach can you use?

Mary Kay Faulkner: Well, I mean, this show that there is some other variation and the results but, I think the wording we used in that article was that overall the results seemed to be favorable for the program, overall, you know, it wasn't that every single comparison showed what we wanted that there was lower maltreatment in the treatment group or high amount of treatment in the no service group but, overall with the design in the number of comparisons we're doing it seemed that there was evidence, sufficient evidence that was favorable for the program. Okay, but, from M. H. perspective again as a statistician and, you know, I mean looking at earlier because this was quite right, this was not quite right, this was not quite right, but, I mean in RCD you have a lot of things that are quite right too and this was just another way to dig in there and see if we could figure out that the results were consistent was what the original findings were. You understand what I say it was like a second round of dealing with this data, working with the data, analyzing the data.

M. H. Clark: So basically all the stuff that I really love to talk about is not what's going to be included in that stakeholder report.

Mary Kay Faulkner: Well, I mean in the article we laid out the, you know, the challenges and the limitations and like the attrition part of it and going in and doing more work on the attrition analysis where the people who were there because there was missing data, were they really different from those who were in because they had data, you know, so looking at the relationship between those two groups and the outcome was their differences. So, I mean it's a laborious process, it's rigorous you got to go through all these different steps, document what you've learned but, you know I like it because it's less invasive, it lets the program function as it should with limited intrusion and I guess

**Session 6.15 – Validity in Quasi-Experimental Designs To Determine Home Visiting Program Effectiveness**

it's just an additional opportunity to confirm results from another, an earlier attempt, or it started this way, and you can do it this way respectively. But, M. H. is the expert I was the one who followed the instructions to complete the analysis. Do you have any other questions? I think we've got cards and, we've got, you know, our contact information on the PowerPoint so you can certainly get back with those and if any of you want more explanation on a particular aspect of this we can, I can provide the way I did it, or how I did it or M. H. has great documents available, references on how to do this with SPSS, which is helpful. So, anything else that you want or are we getting kicked out of here?

Audience: *[Indiscernible]* [01:34:00].