Multilevel models for clustered continuous outcomes, part 1—2010 MCH EPI pre-conference training workshop

SUBHASH ARYAL: Okay, let's start with the – Hi, my name is Subhash Aryal and I'll be doing this training for the Multi-Level Models. Our focus is Applied Multilevel modeling. And for today's work, we'll be looking at, let's go over the course outline. And there is a slight change in this course outline because the time also has been changed a little bit and then I will be doing some part of it and Dr. Hedeker will be doing some part of it, so even though the schedule – we are going to cover everything that's here on the schedule, but only whatever we cover today, it will be a little bit different from whatever it's here in the schedule.

So, the main focus today – the main focus of this workshop is going to be on the two-part – two-part model. One is going to be the cluster data and the other is going to be the longitudinal data. So today I will start with the cluster data for continuous outcomes, Multi-level Models for Clustered Continuous Outcomes. And from this one I am going to go to whatever we have it for tomorrow, which his the Multi-Level Models for Longitudinal Outcomes. So I will do the Cluster Outcomes for Continuous Data and then I'll move to the Multi-Level Models for Longitudinal Data. And then at the end of the day, if we have time, I'll get started on this for the Longitudinal Data for the Binary also. And when Dr. Hedeker comes, he will – then he will start here from the clustered models for binary data and then he will cover the rest of the non-linear model and the same thing for the longitudinal data also.

So the things that we are doing is this one, is the Cluster Data for the Continuous Outcome. This is clustered binary data. And then this is going to be the clustered ordinal and normal data, that will be, that's what we are planning to do today and for the remaining it was longitudinal data which is binary longitudinal data and then this is going to be longitudinal data for when there are – when the missing list cannot be ignored. How do you do with some of those things? These are basically patterned mix of models, so we'll be covering some of these things also.

And most of the work that we are presenting here today it's available in this book by Dr. Hedeker and Gibbons on longitudinal data analysis. This mainly focuses on the longitudinal part of the thing and the examples that we are using here today, some of this data has already been published by Dr. Hedeker with some of his other colleagues and we'll be focusing on those data that is already been analyzed and published in some of those papers.

Now, let's get started, even though it says we are going to start at 9:30, so I don't have any other things for the introduction, so my resume is brief, so the introduction is going to be brief also. So, okay let's go to our training. And the first one we are starting today is the LS, SAS, SBS, SO – where is it? This one. Clustered LS, SAS, SBSS.

The name of the file, so it'd be ClustLS_SAS_SPSS. Now all of this lecture materials after the what's up is over is going to be uploaded to the website and then Dr. Hedeker has a dedicated website that he uses for his class at UIC also, which is UIC.edu under **** Hedeker, and then in there you'll find for multi-level models, longitudinal data analysis and the lecture notes that he uses for the 15-week class that he does at UIC; so all of those materials, the data sets, all of those are there. And we won't be introducing those softwares today, but Hedeker with some of his other colleagues, they have made some free software also called, Mixed Suite, which has like the – it's a free program that's available to do multi-level models for continuous data, ordinal data, binary data, and there are like maybe seven or eight of those things that are available for free. And if you are interested in using that software, you can download that software and go over the lecture notes that's on his website. And the software to use all the data saved that are already there on the website as the Help Files. So whenever I am doing anything on multi-level models I suggest the – suggest that if you do it, it will help you get started with the examples because all the help files and how to run those files are already available over there.

So let's start at what are Multi-Level Data. So we are here today because all of us are interested in dealing with multi-level data. So how does this multi-level data arises? In many cases, like let's take an example in education. One of the things that we'll see is that whenever we are talking about education, we'll see that there are students who are within a classroom. And within a school there are many of those classrooms. And then when we start talking about things at the County level or State level we will see that within a county there are multiple schools and then when we go above that schools level to the county within a state, there are so many different counties. And usually whenever there is a comparison, people will be saying that, okay, this school is good, this school is bad, this county is doing well, this county is not doing well.

So, when we have those kinds of data where things aren't nested within each other, we have a higher level over – one level over here, second level over here, third level over here, multiple levels of nesting. Because of that, we need to take into account the fact that those data are nested. That the students are nested within the classroom that means that all the students in the class – students belong to a classroom, there are multiple of these classrooms. These classrooms are now nested within the school in the sense that all of these classrooms stay the common property of the school. And then when we talk about classes, schools within the county, we say that all these schools are nested within the county saying that all of these schools are within the same county.

And the reason that we are to take care of this hierarchy of nesting depends upon where we are, where we live. Some us live in neighborhoods that would pay very high taxes

and the schools are good. Some of those neighborhoods and some other neighborhoods maybe the taxes are not that high and the general impression is like, oh some of these schools are not good. So when we are analyzing the data from those kind of structures, we are to take into account that, oh we are coming from this place, we are coming from this place.

Even within the school, there may be some – there's always that, oh this professor is good or this teacher is good. So some of the students belonging to this one person, who has been taught by this person and the other students that have been taught by this other person. So all of these here are common property.

So if we do not take into account that common property, then what we are ignoring is the relationship that the higher level of nesting has on the lower level of nestings or basically when we say that, the classroom, all the students within the classroom, maybe they are same. So we have to say that these people are nested, or these people are correlated. If we fail to take into account that fact, then our analysis, it will give us some biased or results that are not really correct.

And that's what we are going to be focusing on today. So data that are hierarchically structured, nested, or clustered. Data collected from units organized or observed within units at a higher levels, so some of the students who are within classrooms, siblings who are within families, and then the other ones is the repeat of those reasons. Sometimes whenever – we may not have people clustered within the same unit, but we maybe observing students, let's say over a period of time. So I'm taking the students measurement at when they are in the seventh grade. I'm taking the same students rating scored when they are in the eighth grade, ninth grade and tenth grade.

So basically what happens is like those four measurements, they are related because whatever is the student's score that is going to be on the tenth grade, whatever was that score in the seventh grade is going to have an effect. If the student is already reading at a very high level while they were in the seventh grade, then they will be reading at a very high level in the tenth grade also. Or if the students are reading at a very really low level, then by the time they are in the tenth grade, their improvement not be that much. So basically, those observations that we have, those are going to be correlated and we have to take care of that correlation.

And in this case, the multi level comes because the subject is the higher level and the repeated observations are the nested thing, so all the observations from the subject are nested within that particular subject student and then we have multiple of those students.

So these are examples of two-level data because we said two level because we just have two levels, the students here are nested within the classroom, the siblings are nested here within the families and the repeated observations are nested within the individuals.

Now when we say Level One measurement of primary outcome important mediating variables. So in the level one, whenever we are measuring those are the primary

outcomes. Suppose if you're studying the reading level, then we are measuring the reading level over here.

And then when we say level two provides context or organizes enough level one units. So basically the multiple observations that we saw from each individual that is the level one data, but the subject itself under which all those observations are nested, those are the level two data. And in this case for a level two, we may have other variables also, so whenever you do any kind of study, you collect a lot of variables, so some of those variables are going to be on the level one and some of these are going to be on level two.

Now what is multiple – multi-level data analysis? When we say multi-level data analysis, we have collected the data where people subjects or units, whatever they are, they are nested within in a certain hierarchy. And if we take into account that hierarchy whenever we are doing our data analysis, that is a multi-level data analysis.

So in this, whenever we are doing something for the school, students within a classroom, then we'll say that if we take into account both the fact that there are students but the students are nested within the classrooms we are basically taking into account both those levels; the students and the classrooms.

Yes.

Question: When you talk about the different levels, it sounds as if you referred to something as a higher level. How do you define what the higher level is because that usually ****.

Higher level. Okay.

Question: If there's a question, can you repeat it so it gets up on the mic.

Okay. So how do you decide what is the higher level. Some of it is just by natural ordering that we get that whenever we are talking about, let's say something like a student, classroom, schools that's the natural way that the data are already set up and then we are taking – whenever we are collecting the data, we are collecting the data at the subject level, which is basically the student, and then the classroom becomes the higher level where the students are there, and then we have the schools where all the classrooms are within the school.

So that's how usually it's the natural ordering when the data is set up well. When you set up your study, it will come into that. When you go to like higher levels of nesting, these days people are doing like four levels of nesting, five levels of nesting. Now if you did the school example, if you talk about a four level of nesting, then it will be like the students were within a classroom, the classrooms are nested within the schools and the schools maybe within the school district or counties.

Question: So then it seems that the higher level is the furthest distance from the individual.

Exactly. Exactly. So, we are – usually we are collecting data at the lowest level that we are interested because whenever we are talking about students, our goal is always to make some inference about the students. How the students are doing, but ultimately when you are talking about the students, even though we are talking about the students we have to take into account the fact that these students are not independent, they are coming from a different level of hierarchy and that's how we define the different levels of nesting.

Now, when we are doing this multi-level data analysis, this is the key thing. Analysis that models the multi-level structure because all of us have done, and like basically whenever we do the first class in statistics or bio statistics it's the Igversion model, right? Whenever we are talking about the Igversion model, one of the basic elements of an Igversion model is that, that observations are independent. Right? We say that, oh why idea follows normal distribution with this mean and this variance and then we say that all of those observations are independent. Now if we ignore this multi-level, that means we are doing the Igversion analysis in the sense that all of these observations are independent.

Now suppose I want to do something with this class, if I am taking some measurement at the – tomorrow whenever this workshop is over, we'll talk about – well all of you will do an evaluation. Now suppose I want to analyze that data for the evaluation for all of these training sessions. So if I just say that, oh there is no effect of the instructor on the satisfaction level that all the students have over here, then basically what I can say is, I can combine all the subjects who are taking in this workshop and say that, okay, we'll just analyze the data as if saying that there is no relationship between the level of satisfaction and the instructor, but that may be doing the thing not correctly or too simplistically.

The better approach would be to say that, oh, there were two instructors for the multi-level model, so all the students who are in the multi-level workshop. That means they are nested within the – this instructor. So whenever we do that analysis like that, were we are saying that all the observations within a particular instructor are related because the way I conduct this workshop it will affect the evaluation that you gave me and it will do the same thing for the other workshops also.

So, if we do the Igversion analysis, simple Igversion analysis, we are saying that there is no relationship, let's just assume all the subjects are independent. But if we do it in the more correct way and we are taking into account that the data are hierarchy or there is a nesting structure, then we are to do the multi-level analysis and that's when we say that recognizes influence of structure **** individual outcome. So when you do an evaluation that means whatever was my – whatever the impression that you get or got about me and about Dr. Hedeker, that will influence the outcome that you present and

that is the multi-level – correct multi-level analysis will take care of that. Whatever was the influence of the structure on the individual outcome?

So when you talk about structure, again, for the two-level nested design, the structure in school classroom and then the students who are within the classroom.

The next one if we talk about is the family. And the family is the siblings. If we are doing our study on siblings, then we'll take care of the fact that those siblings come from the same family. And then whenever we are talking about **** observations, this is the longitudinal case; this always repeated observations are nested within the individual.

Does anybody have any questions? Okay, let's move on then.

Now here's a rule, whenever there is no question, that means either I am doing too good a job and you are understanding everything, or I am doing such a bad job that you are like, oh yeah, we don't care. So, yes.

Question: I have a question. So if the data only collect, say for example, we have the individual level data and also the state level data but we know there could be some intermediate like, level of data, for example, the county level data, but we do not have the data available, so can we just analyze individual level and the state level **** so that's two levels?

Answer: Okay. So the question is that, suppose that the data it's individual level data and those individuals are from a particular state, but within the individual and the state, there is another intermediate level where it's the county level. Right? So we don't have the information on the county level, so how are we going to be analyzing that data for the county level? So do we separately do it for the individual level and the state level, so how do we take care of that?

Usually, the thing that has been done till now has been before this multi-level model became popular, a lot of times the higher levels of nesting they were ignored. This basically, when data was collected, even if it was like nested design, people usually said that we are going to ignore the higher level of nesting and then we are just going to analyze at the simplest level because all of us are very familiar with the Igversion model and we'll just go ahead and use the Igversion model. Now in such cases if there are any kind of like surrogate variables that will allow you to take into account the county level data, you may be able to use that information also. Or if there is no data, it may be a better option to just say that the students are nested within the state and analyze the data taking care of those two levels of hierarchy rather than just saying that, oh I am not going to worry about the state level data or the county level data, school district level data and just do the lower level.

So my suggestion would be, whenever there is any kind of nesting, we should always take care of the fact that this data are not independent because as soon as we take care of

that nesting structure, our results are going to be more valid. There is going to be less bias in the outcome of whatever we are going to get.

Yes.

Question: **** observation do not show any relation in the state?

Answer: Okay. Sometimes – the question is, like what happens when there is no correlation. Right? And we'll see that in the examples that we are doing also that in some cases there may be different levels of nesting, but even if there is nesting, the data may be basically independent of each other. When we see that, then it's okay to analyze the data as if they are independent, but in statistics usually the best option is whenever we are making any kind of ****. It's to go for the **** that is the most restrictive. If we, instead of saying that the data are independent, if there is nesting we are to say that the data are correlated. If the data does not support that, then we can say that it's independent. And then draw an analysis, but before doing analysis and just assuming that the data are independent because it's going to make it easier to analyze our data, that maybe – that's when we run into trouble. If you write a paper with that kind of thing, more than likely the reviewer is going to say that, oh why did you make this ****, what is your rational behind making this ****.

So if there is any kind of **** to be made, unless you have very strong empirical evidence that you get from the data, always go for the restrictive **** saying that this data is correlated. Let's analyze our data as if they are correlated. If the data does not support this correlation **** then we can always go to the simple, simple linear Igversion **** and analyze our data.

Okay, this is nice now. If I'm talking and you are asking questions, then it becomes a two way process and it becomes easier for me and it will be easier for you to understand also. Yes…

Question: Just to understand, to be sure I understand what you said. It almost in most instances, we should be looking for more than one level because it's really rare that one group is not nested within another group. So is that what you are saying?

Answer: Okay. What happens is, it depends – the question was like, should we always be taking care of this nesting. Right? Always when we say that the schools – the students are nested within the classroom, the classrooms are nested within the school, so do you always take into – do you always do this that they are nested within each other. It depends upon the design and whatever questions that you are asking. In some questions it may not matter whether the students, if we treat the students as independent or if the students are nested. So if the design that you are running, if it demands that we take care of the fact that the data are nested, in that one definitely we are to take care of the nesting.

In some cases it may not matter that we will be able to assume that they are independent, especially whenever you are doing some kind of study where you are looking at some

ecological data, or those kinds of things. In those cases, it may not matter for you to do – take into account this nesting. But it depends on like, whenever you are designing your study, whenever you are collecting data, if you're data are going to be influenced by the higher level variable in that case you may have to do this multi-level modeling. If not, then you don't have to worry about it.

Now, why do we do multi-level data analysis means? It assesses the amount of variability due to its level. So whenever we are doing this, basically, in statistics the whole goal is to account for the error. Right? On the unexplained variance. How can we explain most of the unexplained variance in the data? So if we have this multi-level, basically that will allow us to **** this variance into the family level variance and the individual **** subjects nested within the family. And in those cases we have **** in the variance.

Now model level one outcomes in terms of the effects of what is the total variance means. The variance – individual variance is a function of the whatever is the variance due to the individual and whatever is the variance due to the family, which is the nesting structure.

And in some cases it maybe that assess interacts in between the level effects. So we have some data that is collected at the individual level and then we have some data that is collected at the family level or higher level. So in those cases, there will be an interaction going on between those two levels. And if you want to study that, we may want to do this multi-level data analysis. And the main key thing here is, responses are not independent. If the responses are independent, then it does not matter. We don't have to worry about this multi-level thing, but if there is anything that says that it's not a good idea to be thinking of these as independent, then we have to take into account that these individuals within a cluster, there is some kind of influence that is affecting everybody and when that happens we are to do multi-level analysis.

Now for this multi-level analysis, this is the one core data **** when I was a student from Dr. Hedeker and that's basically my life also to work as a bio statistician. What happens is, somebody brings a model. Right? They will say that, okay. This model explains this month's variability the data. Whatever the context mean, what do they do? They will say then this is ****, I can't explain it. Then I go, oh, you did this. You said ****, let me add this one variable and explain some of these things, let me add some of these other variables or these levels of nesting that is going to explain some more variability in the data. And then I improve on that model. But whatever model I present, somebody is going to improve on that and that's why it says that one person's error is another person's career.

So basically whatever they say that they could not explain, I'm going to try to explain it and try to make a career out of it. So. And that's what I am doing. Most of the work – part of the work, part of my thesis was why Dr. Hedeker and his colleagues did it for sample sizes **** for two-level model. In that one, it was just two levels. And in that case, whenever I was working, my advisor said, they did it for two levels; let's improve it

from a two-level to a three-level. So basically we said that hospitals, we said counties, counties are nested within the – hospitals are nested within the counties and then multiple observations from the patients who come to see the hospital. So instead of two level, we went to a three level and I said now, all of these variance that we are seeing in the data, how do you take care of that variance whenever we are doing sample size so basically that – and now I'm trying to put it from three-level to a four-level and four-level to a five-level. So.

Now let's look at the data I structured how it is going to be whenever we are looking at some of these things.

Now, let's say this is cluster one. Cluster one is, let's say, everybody who is here in this multi-level workshop and level two, lets say is in the next door and level three is the next door. Right? So basically, whatever we are doing, this is cluster one so within cluster one everybody who is here today, the subjects one, two and ****, this is the total amount of subjects, so within this cluster all the subjects are nested within this cluster. For level two, for the next door, again, in two subjects and then you will see that all the subjects within this cluster are nested within this cluster. So you can have multiple of this cluster. So one, two and each cluster have one of the key features that you noticed over here is, all of these clusters have different subject size. Right? Life is very easy if all of the subjects have, all of these clusters have, same number of subjects. That's basically what we call the balanced design. Right?

In a balanced design, whatever happens is like – if we are 10 clusters, we will say that all those 10 clusters within those 10 clusters there are 10 people in each of the clusters. So when we have a balanced design like that our method analysis becomes much simpler. Whatever we have done in An Over. Those of you who have already taken classes in An Over, done a lot of work in An Over, there is a repeated measurers An Over analysis that you can do. Whenever some of these things are clustered and all of these clusters are equal sample size, we can analyze the data using that approach, but in many cases, especially whenever we deal with something in public health, we will never get all of these things where it's going to be equal number of people.

If it's a longitudinal study, **** people saying that, why don't you come for this study, I record 100 people by the time I'm through with my study, only 60 people will be there. Forty people they decided that they don't want to come. So we'll have dropouts and we'll have unequal sample sizes. So the reason that there's multi-level models have been popular recently is, one we have good computing power. And with those computing power we can take care of more sophisticated methods that take care of the fact that all of these clusters have unequal sample size.

Now the treatment group, if it's a randomized study, we may say that whether the subject has been randomized to the placebo group or the treatment group if it's a true treatment study. And sometimes what happens is, we may not do this randomizing of the subject level. We may do the randomizing at the center level also, so we may say that if we want to test some kind of **** method… excuse me now. My cell phone – let me put it away.

Again, when we talk about randomizing, it may be at the subject level. Let's say we are looking at these new teaching methods. So what we may do is like we may say that the students within one school, one person is going to get this method of teaching, another classroom is going to be getting another method of teaching. That maybe something, an example of where the randomizing that has been done at the high level or the class level. Right?

Or we may say that it may not be possible to do that within a classroom, but in some other sense, in a hospital when the patients come. Some patients may be given chemotherapist, some patients maybe doing some kind of operation. Right? So whenever that kind of happen, that will give us that treatment group whether the person has been randomized to a placebo or a drug kind of thing and this randomization, it can be at the higher level of clustering or at the subject level also.

So size some of these variables, whatever is the cluster size, now outcome variable usually – our goal is to look at things at the subject level. We always want to – most of the time we are analyzing the data even though there is clustering, we may want to analyze our data at the subject level, how the subjects are doing. Whether this treatment procedure works better, so that whatever we are measuring we'll be measuring those things at the subject level. And some other variables that we don't always take into account when doing our analysis is like, something like gender is, where we are measuring those things at the subject levels. So what is the gender of this person, what is the age, what is the social-economic status because those variables matter whenever we are analyzing the data, so those are subject level variables.

So basically the way the data is collected, we have end clusters where "I" denotes the clusters, "I" goes from one to upper case "N" and within each cluster, where "Z" denotes the number of clusters within each cluster, are many subjects. And the key feature here is, this is NI and not "N". That means it allows us to have unequal number of people in each of the clusters. And that is the key feature. If there is equal number of people in each cluster it becomes much easier. And some of you who are doing your work especially with animals, if it's any kind of animal experiment, usually people don't use multi-level model, and in those cases you'll see that. If you read the literature it's usually the Animal models. And the reason its Animal model is for them, this does not matter because whenever you are doing studies with animals, you know how many animals you are going to randomize to each of the group and how many observations. It's not that the mouse is going to run away from the lab, something like that. So you know that how many observations you are going to get.

But whenever you deal with human, you do these studies and you do everything you can to get them to participate and whenever you talk with them, are they going to come? Yeah, sure, I am going to come next week to your… for the measurement. And next week you are waiting, you are waiting, you call and they don't show. So when that happens, these thing become your **** design is studying that has equal number of subjects, but by the time you are through with your studies, especially when it's a

longitudinal study, it becomes unequal sample size and we are to take care of that fact. So.

Now again, while doing with these things, let's talk about now instead of the cluster design, let's talk about something like a longitudinal design where it's the subject and multiple observations from the subject.

So our level of clustering now is the subject or subject one. And for subject one, we have this observing **** subject for over a period of time. And this is the number of observation from that subject. So we have N1 observations from this first subject and for the second subject, even though when we designed this study our goal was to have N1 one observations from each of the subjects. We say that, we are going to do this longitudinal study for six months and we are going to take the measurements every month. So we'll have month one, month two, month three to get six months, but by the time we are through with our study, somebody may have come for all six time points so I have N1 observation from that person. Some other person may just come for the first two time points. So I have N2 observations. So this is how it's going to be.

And now, whenever we are looking at these variables, there will be variables that sometimes change with time and variables that does not change with time. And one of the things about multi-level models is, it will allow you to distinguish between variables that changes over time plus the variable that does not change over time. And those are the time invariant variables and time varying variables because whenever we are doing the analysis if we take into account this fact, it is going to give us some more information out of it.

Now something like a treatment group. When subject is randomized to either a placebo or a drug, that's a time invariant variable, right. Because whenever they are randomized that's when they are going to be throughout the study. So this is the treatment group.

Gender. Something that does not change over time. Now if it is a short time study and if you are just going to be doing it, let's say for six months and if you say that, oh this person is 23 years of age ****, that's 23 that we are going to be using. So in that case time may be – it is maybe a time invariant variable. But if you are going to be allowing years to change also, it can maybe you want to go to the time varying variable.

Now the outcome. Whenever we say outcome means, whatever outcome we are measuring, that's basically changing over time. Right? And that's why we are doing the longitudinal study. If the outcome did not change over time, we did not need a longitudinal study. So this is changing over time. And something like those. Suppose you are giving different dose of those things and in the example that we are going to be talking about today, we will see that. Whenever we are measuring things that somebody may take a drug and they maybe taking 25 mg. And if we do the measurement after three days, what is the amount of drug that is left in their body, it's changing over time. Right? Every time we do the measurement it's changing. So that is the dose.

And now another very good example that you can think of is like smoking. Whenever we do a smoking cessation study, if you ask subjects "did you smoke within the past week?" They will say, "No." And then the next time they come, did you smoke? Are you still in the program? They will say, "No, I started smoking." And if it's a long study, they will say that – after some time they will say that, "Oh yeah, I stopped smoking again, and then now I am back in the program," those kind of things. So basically, within the period of your study you will see that the variable **** status. It was one time smoking, non-smoking, smoking, non-smoking. So we can take care of that fact that these variables are changing over time.

So again, here in this case, One 2N is the subject and all the observations. This is the longitudinal study, a repeated measure study where we have all of this information. And then **** from One 2N, and again, the key thing over here is there are unequal amount of observations from the subject. And that is a key feature of usually all of us who have done this longitudinal study where going through it, people dropping out and then oh, what are we going to do. Too many people are dropping out of our study. So that is – but the good thing about the multi-level model and the mixed random mode that we are going to be talking today, even if the subjects drop out we can still use the information that we have collected from those subjects. We don't have to throw out the data.

If you are using the repeated measures An Over kind of thing ****, if somebody drops out, means that whole information is gone. So suppose you have a study that goes for six months. This subject came till the fifth month and decided not to come for the last time measurement. So in that case, if you are just using the An Over approach, what happens is you are going to be dropping that subject who has come to the five time point; so then basically losing all the information that you collected. Now with this method, we can take into account those subjects we can use those subjects whenever we are analyzing our data.

Now some of the names that you must have heard of while dealing with these multi-level models is some random fixed model. And in some places, some other people call it, random coefficience model. The other key common name is the mixed effect model or the hierarchical linear mode. So these are different names, but all these represent different – depending upon which field you are on, in some cases somebody will always call it random effect model somebody will call it mixed effect model. But the key feature of all of these models is that there is a random component. That basically we are using some random complement at the subject or clustered level to account for the different levels of clustering.

So useful for analyzing data. What kind of data would we be using our multi-level models for is for cluster data where subjects are nested within the clusters of when we say clinics, hospital, families, work sites, classrooms all of these things where somebody is nested within the higher level of structure and it's longitudinal data. Longitudinal data is the repeated observations of the same person which is – and now sometime what you'll see is that you'll see something where both of those things are employed at the same time. The cluster and the longitudinal. So we may say that all of the subjects within the

schools, the subjects are clustered within the schools, so that basically gives us the two levels of clustering. But for the subject, we may have multiple levels of multiple measurements. So the measurements at the individual level is the longitudinal **** measurements from the subject, but both subjects are nested within the school classroom, whatever it is. So that basically is a clustered and longitudinal and that's a three-level nested design. And when you start increasing those levels of clustering, you may get a four-level, five-level, different levels of clustering.

Any questions?

Question: [Inaudible]

Answer: Neighborhood models. Now, I have not heard the name neighborhood models in the literature that I am familiar with, but it maybe that it's neighborhood models because things are taking correlation because it's correlated data. So for those correlated data, most probably maybe they are being used in some other literature.

Question: How about polynomial [Inaudible]

Answer: No, polynomial, so basically what you do is like instead of having one variable you have multiple variables. Right? Whenever you are doing – let's take an example of Igversion model. It's that you have 'X1', 'X2', 'X3', so basically you have different of these things. Right? Or you may go from like 'X', 'X^2", 'X^4", so different polynomials. So those would be. Yeah, and these models will have different polynomials. It can take, whenever you are dealing with time, you may have "time", "$time^2$", "$time^4$", so you may go that way or you may have different "X1", "X2", "X3", so it maybe polynomial in that sense.

Yes.

Question: How about individual growth curve modeling?

Answer: Okay. Individual growth curve… the question was, individual growth curve modeling. So individual growth curve modeling means, usually whenever we do longitudinal study, that's our interest. Right? We want to see how the subjects are changing over time and that's why it's called the growth curve model also. Growth curve models are more popular whenever you talk about the multi-variant analysis of variance of the **** over they will say that the growth curve model. But here also, we will l do the growth curve model in how things are changing over time because that is the key thing. Right?

Whenever we are doing any kind of study, one of the key goals is not just to see how the group is changing, suppose if we can identify within a group, particular groups cluster or subjects for which the thing works. So in that case, we want to see how the subjects are changing over time and that would be the growth curve modeling.

Yes.

Question: The structural phasing model?

Answer: Structural equation modeling. That is one of the most popular things. It's again, very similar to whatever it is and then sometimes depending upon the literature that we are going to be using ACM. ACM, I won't say it's exactly the same thing because ACM is a broad rubric, it encompasses a lot of things, but yeah, these are part of the structural equation models also.

Now a lot of time, these days you'll here the structural equation modeling, especially if you are in like Switzerland behavioral science. Those things that's called structural equation modeling, but it's basically the same thing.

Yes.

Question: Can you talk about the effect that your sampling design would have on what levels you choose in your nesting? For example, if you have already stratifying your samples to select from schools and then your sorting your data set by those nested by that ****, then is it really necessary to use that as another analysis level and then you select a different type of level within that school.

Answer: Okay, the question was, when you have this multiple levels of nesting, how do we deal with it whether we take into account – that depends upon what your question is, what you are trying to answer. In some questions that you are trying to answer, it may be that whenever you have done all of these levels of nesting you may want to do it. In some cases you may say that, okay, we have all of these different levels of nesting, but some of these levels are not important, some of these are there because you are not interested in it. In those cases, you may be ignoring those also, and part of it is the software that is available also.

In the last – like if you look at the literature lately, maybe like even 15 years ago, a lot of time people ignored the higher level of nesting because there was no software available to take into account that higher level of nesting. But now, with like, even the PC SAS that we are going to be running over here, that's pretty powerful, it can take care into – it can take care of the fact that the data are nested in multiple levels. So when that happens, then we maybe taking into account that fact. And we will see that whenever – in one of the examples that I am going to present how the levels of nesting, if we take care of the nesting and if we do not take care of the nesting how that is going to influence the outcome that we did. We're going to be talking about that.

Yes.

Question: How to determine which variable has, I would say, random effect and which variable has **** --

Answer: Yes. I'm going to talk about that because that is the whole key of this talk. The question was, how are you going to say which one is the random effect, which one is the fixed effect. Right? We are going to talk about that. We'll present the two levels of nesting within subject model and between subject model and that is going to allow you to say that or this one are going to have and this are not going to have… we'll talk about that.

Okay, we are right into it now. When we were talking about – let's now consider a two-level gender model for cluster data. Consider that a model with equal variants and one cross **** also. Basically the way the literature notes have been set up is that we have tried to minimize the math as much as possible because the focus of everybody who is in this workshop is not about doing the math that is behind the mixed effect model, it's like how we apply the mixed effect model and how we interplay it from this mixed effect model. And this is just a notice and this is the standard notice. And if you are writing a paper, this will be a standard model that you will be presenting. So we are just presenting the model here, but we won't be dealing with vectors and matrixes, those kinds of things unless we have to do some kind of analysis in SAS. So this is just a very general model where we are saying that, this is a vector, an observation **** of observations that "Y" denotes the **** of observations. And "X" and Beta, and "X" is the design matrix. Right? "X" is the core various that you have collected. Usually whenever we talk about the design matrix, basically we have what? The intercept, which has **** of one, and then all the observations that you have collected for the subject. So that is here in the XI Beta. And Beta is that – it's the covariates for each of the covariates that is the parameter. Right? Better known as the intercept, Beta1 is the slope, Beta2, all of those things.

Now Epsilon I here is the random effect. So the thing is, these are mixed effect model and we are going to be putting some random effect over here. So what is that random effect? That random effect is denoted by Epsilon I in the note ****. And EI is the error, which is the error vector. So, YI x N x 1 vector of responses for cluster "I" means, if we are talking about a particular classroom, all of the subjects within the classroom. Now XI's, NI x P x X1 covariant matrix means, NI is the number of observation, P+1. "P" is the co-data. Suppose we collected data on, let's say gender and ****, so "P" is 2, but whenever we are doing our modeling we always put an intercept in it. So that is right. It's P +1, **** 2 +1, that's why it's **** for the design matrix.

Now, "B" is the B +1 x 1 vector of regress and coefficient. So Beta not is the intercept and let's say if it's gender and ****, Beta1 and Beta2 are going to be we are talking about the coefficience for gender and ****. So that's why it's B+1.

And this other cluster effects so now, these are the random effects for the cluster to take care of the fact that this data within the cluster are nested. So now those are the cluster effects. And whenever we are talking about a random quantity, we have to make some adjustments about that, what distribution it follows. And here we are saying that this NID means, normally and independently distributed, with mean zero and variance **** Epsilon. So basically when you first did a class in statistics, the one that we deal very

frequently is the normal distribution. And the normal distribution is basically characterized by two parameters. One is the mean parameter and one is the variance parameter. And we are saying that the random **** follows normal distribution with this mean and this variance.

And then EI. EI and NI x 1 vector of residues those are basically the error. And we said that the errors are NID with this mean and this variance. So basically what this means with this structure where you say $Sigma^2$ I NI means, once we put a random effect in our model, the basic premise of the basic **** of this model is the data are correlated. By putting a random effect, we are taking care of the correlation. And once we take care of the correlation in the data, once the random effect is there, the data are basically independent. So once you put the random effect in the model, now you can assume that the observations are independent. So basically we'll say that conditional on the random effect given that the random effect is in your model, the data are independent and that is denoted by this ****, if you look at the matrix you see that this basically means that this data are independent.

So the key thing that you need to focus while you are looking at this model is, given the random effects, once we have explicit random effect in our model, our error becomes independent. If we don't have this random effect, if we forget about this random effect, this basically is a regressing model. Right? In a regressing model we have the design matrix and the parameter and the error. In that case, whenever we are doing the Igversion analysis, we are already assuming that these errors are independent. But now since our data is clustered data and the subjects within the cluster are nested, to take care of that fact, we put in a random effect and given the random effect, now the observations are independent of each other.

Question: Question please.

Answer: Yes.

Question: [Inaudible].

Answer: You mean this Beta? Yeah. The beta is the vector of **** --

Question: Also the only one would be – before the Epsilon. ****?

Answer: Yeah.

Question: The other one, other one, ****. The cluster effect.

Answer: This cluster effect is the random effect.

Question: Well **** the notation should be like a subvector, eh?

Answer: Yes.

Question: You are in like a two vector, eh?

Answer: Yes, basically this is – it has a **** "I", so basically with this **** "I" what it is saying is, for each subject or each cluster, there is a random effect, so that's what its taking care of over here with Epsilon I. So for each subject we have a different random effect. So that's what this is "I" is for. This is for the individual. "YI" is the individual. So for each individual, this is the response. This is the design thing. This is the random fit, and this is the error. So that's why we put it like this.

Question: When you say that huge random effect, do you mean that that effect **** after randomizing ****?

Answer: No. We are talking, we are going to interpret, we are going to talk about what does this random effect means, how it is taken care of. We'll talk about in the next few slides. Yeah, that – I am going to focus on that one.

Now, as clusters ****, "I's" represent, "Y" and "X" cluster sample size can ****. So that is the key. Right? We have an "I" **** for each subject, it can have different model observations and different for the covariant also. The covariant matrix "X" can include co-variants measured at the subject level, co-variants measured at the cluster level and **** level ****. We talked about some of the variables that are measured at the cluster level, some of the variables that are measured at the subject level and there is a cross level interaction.

So this "X", which is basically the design matrix that we have dealt with in all the Igversion class, everything though that is the design matrix. The total number of co-variants determines how many variables that we have studied. Those are the co-variants.

The number of columns "X" is people **** want to include the intercept. So usually whenever we are interpreting the intercept is the value of "Y" when "X" is ****, all of the "X's" are zero. Right? Are the co-variants are zero.

In many cases it may not make any sense to say, oh when we were talking about ****, if it's a simple model so that this is the value of "Y" if the **** is zero. I may not make sense to be just interpreting it at that, but we need the intercept. If now, we are always assuming that the value of "Y" is zero when "X" is **** to zero, and that may not be a very reasonable ****. So we'll always put an intercept in our model so that we can look at it more correctly and in some cases intercept will have its own meaning. In some cases the intercept won't have any meaning and some of you may have already done this analysis to make the intercept more meaningful. You have centered it at the mean or something like that. But in this case, we are just going to be putting **** intercept – all of our models is going to have an intercept. We are not going to do any models where there are no intercepts.

Now, these are the random effects. Random parameters distributed with this distribution distinguishes model from the fixed affect that multiple regressent models. So how is our model different from the regressant model that we have? We are have explicitly put a random effect over here. And what does this random effect represent? This random effect represents the subject cluster and one for every cluster. So basically we are put a different random effect for each subject and that to account for the fact that subjects are clustered within the – in those particular clusters.

Now if subject clustering has little effect, estimates are zero, variance is zero. So basically if there is no relationship, there is no cluster effect, what is going to happen on the estimate of the cluster effect? The estimate of the cluster effect is going to be zero or basically we say that there is no variance around the cluster. Right? Everybody is at the same point. So we don't have to worry about the people being different.

Now if the subject – if the clusters are being different – if subject clustering has effectiveness, if there is a cluster effective, things are different between the cluster, then the estimate of this cluster, which is the estimate of the random effect is going to be different from the zero and the variance will increase. Basically what happens here, if there is no variance in the data we are talking about a point? Right? But if there is variance in the data, it's going to be more spread out. And we will see some of these things graphically and it will make more sense when we look at graphically also.

Now ultimately our goal is to model the "Y". Now what is the distribution of this "Y" is? "Y" is normally an independently distributed with this mean and this variance. Right? Basically you have seen this in the regression model analysis also. But we have multiple observations, so it looks a little bit complicated over here, so these are matrixes, but they basically this is the assumption that we are making about the observations. That the observations have this mean in the population, this is the mean of the observations and this is the variance of the observations.

Now usually it means from multiple Igversion model, so whenever we are talking about this is the mean, that's what you get modeling. What is the average effect? If I look at something, a lot of the times it's not the individual that we are interested on, as a group, as a whole, what is going to be the mean? And that is what we are going to be looking. The variance, co-variance structure accounts for the clustering. So basically by having this variance, co-variance over here, we are taking care of the fact that the data are clustered.

Now how do we get it? We don't have to worry about all of these matrixes, we'll be using SAS or SPSS, whatever software you are using and that is going to give you all of this variance, co-variance estimate. Our goal is to look at this variance, co-variance estimate and see whether it is significant or not. Do we need this variance, co-variance in our model or not.

Now for the variance – yes.

Question: Can we go back to variance?

Answer: Okay, sure. Sure.

Question: When you talk about the subject clustering has little effect. Does it mean the subject's – we're seeing the cluster are not correlated into ****.

Answer: Yes. Now, the question was how does subject clustering has little significance. If we are saying that this clustering that when we are taking into account that the observations are clustered, if we are to take into account this cluster **** variance, how are we going to know if the cluster is significant or not. Right? Usually if the cluster effect is not significant it is going to have an estimate, the cluster estimate for the cluster variance that is going to be very close to zero. If there is not much variance means within the cluster means there are very similar we may not have to be taking into account all of these different clusters.

So everybody, suppose I put a model with clusters for… a random cluster effect for all the three groups and then if the effect of the cluster is very close to zero and it's basically what I can say that or I don't have to worry about the clustering effect. I can take care of all the subjects as similar and then just do a simple regression analysis kind of thing.

Now if we go to the clustering effect. If there is a – if the outcome that we are going to be getting from this group versus that group, if the **** the other group, if it's different, whenever we model the cluster effect usually that there is much variance in the cluster ****, the groups are different from each other. So in that case when groups are different like that, all the subjects within the group are similar, but the groups are different means you have to take that into account that cluster effect.

Question: So **** if we have this date as the **** tracking outcome, it can – where you live. If you live in State A, it will be different from –

Answer: State B. Okay. The question was, if suppose, let's say the state was the level of clustering, how are you going to be ****? If the states are very similar to each other, right, in whatever we are measuring, in that case the cluster effect is going to give us a variance, a random effect for the state, it going to be very close to zero and the estimate is going to be close to zero. So in that case, we may not have to take into account the fact that the subjects are nested within the cluster, within the state. We may say that we'll just take care of these as 1,500 subjects, 1,500 subjects. Now if there is a clustering effect, if things are – if the state is influencing the outcome ends, we are going to have a random effect for the cluster and if the variance, if there is a lot of variance within those clustered means, we have to take care of the fact that these are nested. Okay.

Now variance correlated and structure accounts for clustering. Within the cluster, this is the variance and this is the covariance. So basically this is the variance in the data and this is the covariance. Now, more than covariance – all of us are very much comfortable with the term "correlation," right. And this is basically the correlation. Covariance, there

is a direct relationship between correlation and covariance. So you can think of this as a correlation also.

Now, this is called a Compound Symmetry Structure. What is meant by compound symmetry structure is, basically we are saying that the variance over the different cluster is the same and the correlation between the cluster is the same also. So, if you are familiar with literature of some – if you took a lot of classes, the statistics classes that you took. If it's from a Stat Department you will see that they will talk about all of these models called the autoregressive model,
Add More Model, all of those different things. It's the same name, but basically what we are saying with this compound symmetry structure is, we are saying that the variance, covariance structure is following this particular pattern.

Now in all of our work, we may not have to make that **** that all of the – every time you **** followed the instructor. For our work we are just going to be dealing with this model which is called the Compound Symmetry Model means all the variances are the same and all the co-variances are the same, but there may be some general model. If you do some **** work you will see that just this **** may not be varied and you may have to go to a little bit more complex thing also, but for right now, for cluster data, this is a very reasonable ****. So we will not really dealing with the compound symmetry structure and next when we go to the longitudinal data, this compound symmetry structure is not always reasonable. So we may talk about some other structures also.

Now the ratio of the cluster variance to the total variance is the intraclass correlation. Right? Every time we do any kind of sample size calculation or whenever we are reading any paper of those kinds of things, usually we will see this term, "intraclass correlation." So what is this intraclass correlation? Intraclass correlation is basically the ratio between those two **** out of the total variance, how much is attributable to the cluster. So that is the intraclass correlation.

So this is the total variance, this is the error variance, this is the cluster variance. Out of this total variance, how much is the variance that is attributable to the cluster and we will see that with different values denoted and in a couple of the next slides you will see that. What are some of those reasonable values, what people have noted in the literature for some of this intraclass correlation.

Now this intraclass correlation, this is the "R" is the ratio between the cluster level variance and that total. Class is a background since education class has meaning so instead of calling intraclass correlation; they are talking these days about the intra-unit correlation. So basically replacing unit with the appropriate interest **** correlation, intra-clinic correlation, intra – but a lot of times whenever you are doing that calculation in softwares and those kind of things it will just say **** intra-class correlation. So **** and values between zero and one. So that's a correlation value, right. So it will take a value ratio of between two variances, so it will take a value between zero and one. So when it's going to be zero, if the cluster level variance is zero, so that means it's going to

be zero and when it's going to be one means your model accounts for all the variance in the data. In that case, this part is basically zero.

If that part is zero, then basically you Sigma squared epsilon by sigma squared epsilon that is going to give you a **** of one.

Now what does it denote? Why are we interested in it? It measures the degree of similarity of measurements within the cluster. How much similar between that within a cluster. With an **** cluster how much other measurements similar to each other or the ratio of the variable attributed to cluster **** direct, this is the variance for the cluster, this is the total variability. So out of those total variability, how much of that variability is accounted for by the cluster? That is the intra-class correlation of intra-unit correlation.

Another way of explaining this is proportion of total. Right? This is the total is the unexplained, this is the variance. Out of this total variance, how much is accounted for by the cluster.

Now **** for smaller clusters. So if there are smaller clusters, which basically that will happen. Right? If it's a very small cluster with maybe less, we are doing it for a student of sample size five; in that case we might see a high cluster. Right? Because maybe the students are very similar to each other. If we take a class of, let's say, 100 students, there is going to be much more variability in the data. The students look much more different from each other, so in that case it may change because of that. So, it's larger for smaller cluster means the subjects within that cluster are very similar to each other. If it's a large cluster means, there is always, even when we are talking about a very large group of people there is a lot more variability in that data, they are very different from each other and that's why usually when it's a small cluster that's used to be ****, and these are some of the values that have been noted in the literature for of the studies that has been done, this is the correlation for the spousal pairs, this is for the practices and this for looking at whenever we are looking at different counties.

Now why these other numbers are important? It's like suppose you are doing any kind of designing or study and you want to do sample size calculation. So every time you want to go do a sample size calculation to see what sample do you need for this study, if you go and talk to a statistician or a biostatistician, they are going to ask, what is the variance? Bring me the variance? Bring me the intra-class correlation?

Sometimes you may have some pile of data from that pile of data you might say that, oh, this is my pile of data. I got this intra-class correlation, use this. But sometimes you don't even have pile of data, so in that case you are going to report to the ****, so suppose you are willing to be doing this study for this spousal pair, then we may be able to say that, okay, some of this are **** in the previous studies they have noted this intra-class correlation, so we may be able to do our analysis with this intra-class correlation.

Now what will happen to this intra-class correlation? The key thing over here is we may be doing the study in the same group of people and lets say we are from the same group of people, we have noted like maybe three are different, four different outcome variables and depending upon what outcome variable that you are studying, this intra-class correlation may change. You may get a separate value for intra-class cor – it's the same group of people, same study, but depending upon the different variables that you are studying and when you calculate the this intra-class correlation, you may get a different value.

And this is one of the example from the study with this problem of Antarctic shell was studied and in this case this is a 63 sisters within 26 families, they were looking at this cluster data. And when they were looking at this cluster data, you will see that in this case one, height was one of the factors, the sight factor was another thing, BMI was the other variables. So depending upon the different variables that we are studying, it's giving us a different ICC now. To calculate this ICC at .487, how are we getting this .487? Let's look at it. So let's go to the calculator to see if it's here.

Calculator, where is it? Okay. So this is the family variance, this is the residue of the variance. Right? So we want to calculate the ICC for the family variance, intra-family correlation. So in this case, if you are looking at it, it's 2.743, class 2.895 get's me 5.638. So it's 2.743 divided by 5.638, is going to give me a intra-class correlation of .486, which is .487. So that's how that intra-class correlation has been calculated.

So if you run your software, whenever you are doing your analysis, this family level variance it will come from your – the software is going to give you this, the software is going to give you that. And out of those two things, basically you need to calculate this intra-family correlation, what you are doing, this is the total variance. You add those two terms and once you add those two terms you put the family variance, which is the cluster level variance. Right? If you put that cluster level variance, if you add those two and if you divide this by the sum of those two, you are going to get the intra-family correlation.

Now in the intra-family correlation for the site factor, this is it and for the BMI, this is the intra-class correlation. And now for this data, these are some of the descriptive studies since this was the main height, this was the variance for the site factor and the BMI. So does anybody have any question about how to calculate this intra-class correlation? Because you come up with this whenever we are reading papers, literature, every time we come up with this thing. So what is the intra-class correlation? Basically it is the ratio between the total variability and the amount of variability that is accounted for by the model, the cluster and in that case how are you going to calculate it? We add the total variance and whatever is the cluster level variance, we divide the cluster variance by the total variance and that is going to give us the ICC.

Now if you want to try this one, you can certainly do it for this one and you should opt in this number.

Yes.

Question: so is it also saying that height is more correlated between the sisters –

Answer: Yes, exactly. Yes.

Question: Than ecological or BMI?

Answer: Yes. Yes. So in this case, the people, the sisters who are in the same group, they are more correlated because the cluster of the sister in cluster is taking into account almost 40% -- 48% of the variability that is not accounted by the other variables in the model.

Does anybody have any questions? Yes.

Question: ****, so that means that a substantial proportion of the variability in height is now accounted for that would have just been the error –

Answer: Exactly. Exactly.

So if we have said that the sisters are independent of each other, yeah, we can do this too study like that, but basically to say that the sisters are independent, especially for identity factor like height, we have to say they are correlated so in that case, and we saw from the data also that whatever was the variability that you are seeing in the height, it's because that they are correlated with these other that is accounted for by the fact that they are sisters and not independent subjects. And we must say that the site factor BMI, they may not be more correlated and then when you are doing your analysis, you will see these numbers, how to make sense of these number means that it correspond to the fact that we think is correct. And in this case you certainly makes sense if you are talking about a factor like height to say that the sister's height are relative to each other is a very reasonable ****.

Yes.

Question: So if you do, based on the BMI, you could say use regular regressions as –

Answer: Yes, in this case, exactly. Whatever – in this case, the family variance, right? The cluster level variance is very – oh, the question was, if we are doing this thing for BMI, why do we do it ****.

Now, basically just from this, we don't want to just go into the regression because we want to do something we'll talk about what is called the likelihood ratio test to see that if this BMI factor random **** from the BMI significant or not. Sometimes we will see very small amounts but statistically they will be significant. We do a likelihood ratio test and see whether it's significant or not and if it's not significant, we go to the Igversion, but more than likely we will find something like that.

Question: So you can **** determine from ICC the relationship –

Answer: No, you just don't – yeah exactly. We don't – every time whenever we are doing this statistical work, when we look at the graphs, these will give us some ideas about how our data is behaving, what we want to do a statistical test to see if that is significant or not. If it's significant, we can say that, okay we have to do it, if not, we don't have to do it.

Anything else? Yes.

Question: I don't understand why there's no or why the family correlation is zero or is so low. I mean, I understand that the ratio how you described, but why in this study –

Answer: NO, in this study, it maybe part **** of the study also. If you do this same study with another group of subjects, right, you may see that they are very highly correlated, but in this particular study we see that those things are, when you go, whenever we are talking about the height for the same subjects, we see a high correlation. And for when you were talking about the BMI, we are seeing very – it's just the factor of the data. Now if we talk with somebody who is an expert in this field, they maybe able to give a more better – but statistically for me when I am looking at these numbers, it's just the factor of the data, that's what I am seeing.

Question: Because I guess some of the sisters were anorexic so the weight and some –

Answer: Could be, could be, yes, yes.

Question: I have a question. How did you get ICC of **** up .01 percent ****?

Answer: Yes, now some, the –

Question: ****

Answer: Okay, the question is that, sometimes we see a very small number, right? And then it's very significant. That's when it comes into you as a researcher, that's what you have to decide. If you are a physician, it may be something that is statistically very significant, but clinically, it is useless. So in that case, if it's statistically significant and clinically meaningless there is no point in doing statistics in that sense and in some cases it may be that you are showing something that is not statistically significant, but clinically it may be meaningful. If you observed that if the difference between the two groups is like maybe six percent, seven percent. And when you do the statistics, its not statistically significant, but clinically to find a seven percent difference between two groups if it's a big ****, yeah, you will be using your own judgment as expressing the **** and saying that, okay, even if it's not significant I am going to put that in the model and I'm going to use it as significant even if it is sometimes statistically significant but clinically not meaningful, it may not be. In public health cases you will see that. Whenever, if you can reduce the disease brought in by maybe one, two or three persons,

whenever you are talking about the population, it may be a huge decrease, but statistically if we are talking about a three percent decrease, the model says that it's not significant. So in that case we have to go beyond what the statistics is saying and rely on our own judgment as researchers and expert in the field in that case.

Question: [Inaudible]

Answer: Okay. Yeah, the question was like, if the ICC is not significant, so if we put some of this random effect is going to give us a very complex model. If we don't put that **** random effect it may give us a very simple model. And that is one of the things we always do while – when we are doing statistics like. It's always easier to interpret the model that has one variable compared to that of the two variable and three variables. So whenever we are doing some of these things, sometimes what we need to do is like we need to use statistics as a tool. It's not that when something is less than .05 – oh yeah, this is less than .05… it happened with me once. I was working with a physician and then he got a data set with 35,000 subjects. And we were doing this analysis and then the **** .00001, he was really excited he said, "Oh yeah. I got a significant result." And then we looked at the "I^2" value in the Igversion analysis, it's basically three percent and I told him, "Your model is not working very well, this is – it's only three percent." And he got really anger at me saying, "What are you talking about? When the "P" value is so low, it's significant." I said, "Yeah, it's significant but when you're model is explaining only three percent variability in the data, then statistically it meaningful, but clinically whenever you are trying to publish it, people are not gong to buy it saying that your model explains three percent of the variability and you are saying it is a function the large sample size."

So we should not take the thing just at the statistical point to a five level and use our own judgment to see whether a simpler model makes sense or more complex models, and then depending upon that decide what we want to do.

Okay, now we are going to be talking about the standard Igversion model for the cluster data. This is a smoking prevention research study that Dr. Hedeker did with some of his colleagues and in the study, what they did was like it was the television school and family smoke prevention and cessation study **** till 1988, it's published under Journal of Consulting Psychology, I think, the Consulting and clinical psychology.

So this data – it's a soft sample of the project were chosen with these characteristics. It was a much larger study out of those, this is why they chose the sample was 1,600 seventh graders and the 1,600 seventh graders were nested within 125 classrooms in 28 Los Angeles area schools.

Now, between one to 13 classrooms per school, so within each school there were between one to 13 classrooms and within each classroom there were two to 28 students per classroom. Now, outcome variable was the knowledge of the effects of tobacco use and basically they wanted to see that the timing was student tested at three and four

intervention so they were measured before the intervention was given and then they were measured at once the intervention was given. So basically it was a two time point study.

Now at the simplest case whenever we are talking about the two time point study and think of it like a **** teachers, right, like whatever we have been doing in the many of those simple analysis is **** **** before/after kind of thing, so this is basically that one.

But in this case the complexity of this thing arises because we have different level of nestings. Right? The students are nested within the classroom, the classrooms are nested within the school, so just doing the tedious and the individual level comparing the students with each other, the pre and post **** the students may not be a very reasonable way of analyzing this data. And basically it consisted of like a questionnaire with seven or seven questions I think, and they scored those seven items and whatever was the total score that each student got, that was the outcome variable that was the knowledge about the effects of tobacco use.

And then for the design, there were like basically four of these four treatment arms, one was the social resistance class curriculum so in this case it was the class curriculum included this thing. And the second one was just like it did not get anything in the classroom I think, but they watched some video on the TV intervention, since it's combined with TV means, both of those combinations given to the students and no treatment controlled group means, this group of students that did not go and get any kind of this intervention.

Now these are some of the descriptive statistics for different groups. Now for the no curriculum and no TV, so this one was the mean and this is the standard deviation. Curriculum with the mean, there was no curriculum but they just got the TV intervention. This was their pre-testing, this was their knowledge score about the effects of tobacco use and this is the standard deviation. And in the curriculum group without any TV, this was the mean and the standard deviation and this was the mean for the group which got both of those that **** scored here. Now when you look at the full **** analysis you see that there's no CC, no TV, there basically increased by a little bit. Right? Their pre-test score was 2.1, their post test score was 2.3, so basically with the, if one group did not get any kind of intervention with time, basically their score of knowledge about effects of tobacco use increased slightly. Now the students who got the TV intervention, we see that their score increased quit a bit. Right? Almost half a point. Before it was 2.08, now it's 2.53 after the intervention, so it's 4.452.

Now for those who got the classroom curriculum that they were told **** students got something in the class but they did not get any TV effect, there was almost a one point increase based compared to the before thing. It was 2.05 and 2.96 and for those who got it – who got both the curriculum and the TV is seems like they got the – the score increased by almost 4.85 points, almost one point and of all of this data, we see that the largest increase is between the, for those students who got both the currr—who got both the in class lecturing and no, nothing on the TV. So these are so basically our now goal in a sense is to compare to see if those are statistically significant. Right? Are these

school significantly different, are the students who did not get any intervention… the students who got some form of intervention are they really doing well? Does this intervention help to make the students understand better about the effect of tobacco use?

Okay, now we are going to the model of how are we going to be putting all of these things in our different models. We are going to be using the – notice that this is a popular way of noticing, it's usually done in education that we are dividing the things into two parts. Something we call the Within Cluster Model and the Between Cluster Model. So what is Within Cluster Model? We are basically modeling what is the score for the individual. The Within Cluster Model gets you the… whatever is for the individual.

And basically what we are doing in the Between Cluster Model is, this individual score, we basically can think of it as a group score and how much this is a group score and we put a random effect over here.

So basically what we can say is whatever score I am getting, right? If we are doing it here in this class, if it's a tobacco cessation, smoking cessation program, whatever score I'm getting, this is my score. Now this, my score, what I can do is I can explain my score as a function of what? Whatever is the group score that we have and whatever is my individual effect depending upon whether my individual effect is positive or negative, I may be – my score may be higher than the group score, or my score may be lower than the group score. And that's what we are doing here.

**** "I" is that intercept. Basically this is the intercept **** individual level, so what is the intercept, this first score, what is the score… my outcome score after the intervention is, my outcome score after the intervention is whatever was my initial score, my the previous score, before the intervention, a function of the before intervention and some kind of error, so basically what you are seeing here is, my first score is a function of my previous score. So if I'm – if the first subject is like seeing that smoking was bad, I don't want to smoke, there is – that is going to influence what is going to happen to my score at the end of the study. And this was derided as the intercept and the effect of the pre-score.

Now how are we going to be **** with breaking down this individual score. The individual score is the function of the group. As a group whatever was the average, whatever I got the previous score… whatever I got as a function of the intervention that I got. Right? The intervention is going to have an effect on whatever is my final score going to be. So that is that intervention effect. And this is the individual cluster level. How much different a cluster is. If it's the individual study how much the individual is different? It it's a clustered study means, how much is the cluster? This is the average score. How much is each school or classroom different from the average score? This is the average score and this is the random effect and this random effect is saying how much is different from the average score.

Now in this case, how do we decide whether the cluster is random or not. Suppose I'm doing my study and if I want my study to be valued for just the people that I've collected

my data on. If I collected data on 28 schools and if I just want my results to be valid for those 28 schools, then I do not have to put in a random effect. Right? But the thing is, even though I do my study with these 28 schools in L.A., I want to be able to generalize my study throughout the country. Right? Maybe not throughout the country, but throughout the county or throughout the state, whatever it is. If I want to make my study to generalize ****, I have to put that into the random effect. The cluster effect is going to be random. But the other thing this random cluster effect is going to do is it is going to say that observations are correlated with each other. This classroom within the schools are correlated, and the subjects within the classroom are correlated. That is also taken care by this random effect.

So, by putting the random effect at the cluster level what we are basically doing is we are basic… here we are not putting any kind of random effect for any of these other variables. We are only putting the random effect at the cluster level random intercept. So whenever we see this random intercept means I have this line. Right? Suppose this is my line about how I am doing. This is my trajectory. Now, this trajectory if it's for a group, now I'm going to be how much farther or lower from that group. So let's say this is my score and the group score is over here. So how much is that difference between my individual score and my group score? So basically that is going to be given by that random intercept. So when we see a random intercept model means, we are saying that all the people whenever the difference between the people that we observe its **** **** **** intercept level.

When we look at change over time, we are basically saying that we are changing at that same rate over time. What are the **** all of the schools are different. Now some of you may say that is it a reasonable **** to say that everybody is changing at the same rate? We can definitely improve on this model and we are going to talk about this, how to improve on this model from a random intercept to a random **** in the longitudinal data ****.

But here, basically we are saying that, and this is a reasonable **** four-cluster design. We are saying that at the beginning level, how much is the different between the subjects. This is the group level, how much higher or how much lower, you can look upon the estimate of these things, is it individual school or individual classroom, whatever it is from the thing. And then basically this Beta1 is the average score for the ****. So how we interpret this Beta1, we have all done it for a **** thing. Right? This is the change in the first score for one unit increase of one unit increasing the pre score. So if the pre-score is higher, my first score may be higher if this is positive. If my pre-score is higher and this is negative means, my first score is going to be lower.

Any questions. Now I'm getting a lot blank looks. So.

Yes.

Question: Well done, you just did that really fast. So.

Answer: Okay. Let me go back and –

Question: Go back and go a little slower.

Answer: Okay, let me go ahead. I can see that, everybody is like, what is he talking about? So let me go back.

So basically, in this study, the goal was to see that what is the effect of intervention on the education level of the student after the intervention is given. Right? We have four different intervention arms. The curriculum, where the student got some kind of lecture in class, the next one was TV. No curriculum was given, no in-class intervention, but they watched something on video. The third one was where they got a combination of both, both the curriculum and the TV and the final was the control group where the students did not get any kind of intervention.

Now, our goal is to model their first score. What is the intervention score after that – after what is the score after the intervention? And basically what we are seeing is, the intervention score – the first intervention, once the students get their intervention, whatever is their score, it is a function of their previous score. So basically, and that happens in many cases. Right?

If we are… if the kids are really like, oh they don't want to smoke means; they might not smoke in the future also. But if they are thinking smoking is cool, smoking is good, then they are more inclined to smoke. So basically, the first score is a function of that pre-score.

Now in this case, what we are saying, what is the first score for the individual. The first score for the individual is whatever was the individual's score that they have intercept at whatever is the level of smoking **** that they had and whatever is the change. Whatever is the change in their initially, how is that going to effect my first score and the error ****. So basically, the first score is a function of that. This is called the Within Subject Model because we are talking now about a individual. You see that we are talking about the individual cluster. So whatever as the effect of the cluster, whatever is the initial value for the cluster, whatever is the change in the post score for the cluster because that is from the individual cluster, and that error.

Now, whenever we talk about the score for the individual, one way of thinking about the score for the individual is whenever we are talking about the group is that this is the group score and how much higher or lower is the individual compared to the group. Right? If I am evaluating myself, 10 students. If I am evaluating 10 students in a class, I **** that the average score for the students in my class is 85, but not everybody is going to score 85. You've got maybe some students who is going to score 100 and then maybe some student who is going to score a 40. Right?

So in that case, if I say that 85 is the average score, that is the average score for the group. Now for each individual, by putting the random effect, what we are basically

saying is, how much higher or lower compared to the group effect because in that case, it becomes easier for us to talk about. This is the average effect for the group. And compared to the group, this person is maybe doing well and this person may be doing worse. Right? If they are higher than the group, then maybe they are doing well. If the higher score is better and they are doing low, this lower score maybe better.

Now this individual score, what we are saying, this individual score is a function of the population average and this is the intervention that they got. Right? Because if our intervention is working, the post-score should be defined by this thing called the Intervention. Right? They have some notice… notion about what is how good is smoking. That is the pre-score. But the reason we are doing this thing is, can we increase our post-score based upon the intervention that we had given. So for each individual this is the initial population level thing, this is the effect on the group because of the effect of the curriculum that is because of the effect of the intervention that we have and this is the random effect that determines how much higher or lower is the subject compared to the group thing.

Now, next to what we have is the Beta1 ****. What is the individual change in the individual? This is the intercept. This is the slope. Right? What is the change in the slope? The change in the slope is the function of this and this is the average slope. So basically how do we interpret this average slope? We say that this is the change in the first score, the interpretation of this one is, this is the change in the first score for one unit change in the pre-score. Right? So, if this has a positive sign, when we estimate our model, if this has a positive sign means, when the pre-score goes up, the first score also goes up. If this has a negative sign, when the pre-score goes up, the first score goes down. It's basically in the model. If this is the "X" axis, this is the "Y" axis. If the line is going like this means when "X" increases, "Y" also increases. And in this case, if a line is going like this means, this beta is going to be negative. Right? If this is going to be coming down means, when "X" increases, "Y" decreases. And this is the "Y", this is the "X". And that's how we'll be looking at it.

Now for the residues, because it's a nested design, we have this thing called a level 1 residue and level 2 residue. At level 1, because this has **** the **** script at the cluster level, **** soft script in this case those are the level 1 because it has both soft script over her and this is the level 2 because it only – there is between the cluster and these are the – and as since we have already saw in the previous slides that these are the **** that we have made about our scores.

Now for the same study, these are some of the parameter estimates that were opting while analyzing this data. Now the data was analyzed in two different ways. Where it was line one time the level of the school was enrolled. So the students who have said that the student are registered within the classroom only and in another way of analyzing the same data, the student school where the classroom level were ignored and then the students were nested within each school. So in this case, this was the intercept. So what is the intercept? In this case there was no pre-score, no classroom curriculum. So this is basically what you would call a no model. Right? Without any ****. And in software

**** that will present a result that says the non-model. When it says, “non-model” means no correlated. We are two **** basically in our model, Right? If you go right here. One was the pre-score, one was the curriculum. Now this model does not have either one of those things. So in that case, this was the, intercept, this was the class level variance, this was the residual variance and this is the ICC. And to find out now, basically what we want to do while doing this kind of analysis is, we add in the next case, we add this pre-test score. Right? Does adding pre-test score to our model that does not add anything, does it make a significant difference? Are the pre-test scores a significant predictor. We are to do a test. Right? And to do that test, we need this thing called a **** value, and that’s what we are going to be using to see – to do the thing.

Now in the next model that is run for the students within classroom, the pre-test score was added as a co-variant. Now let’s see if we can explain more variability in the data by adding the pre-test score. When we add the pretest score in this case, this was not – value for the pre-test score. So what does it mean to get a .320, which is positive means? Every time the pre-test score increased by one unit, the first test score increased by .3 value. Right? That was the increasing pre-test/post-test score for one unit increasing post-test. So in this case, if they are the higher pre-test means, the post-test is also going to be higher and we see that by putting this variable over here, our cluster level radiance has decreased a little bit and then our residue variance has also decreased a little bit.

Now we are again calculating ****. Now, what is this 129.0? This 129.0 is basically what we call the **** square value to compare this basically tells us, is adding pre-test score to a model that does not have anything to it, does it make a significant difference?

Now how do we get this 129.0? Basically, you multiply both of these **** value by 2. Why don’t we try **** here. Let’s multiply these two numbers by minus two. And let’s see. 2760.9 times 2 gives me. So let’s write it down, 5521.8. And let’s multiply this number by also. 2696.4 times 2 gives me 5392.8. So what was the previous number? 5392.8 gives me 129. So basically, this 129, and why do we have to multiply this? You will see that in some software, the software will automatically produce the minus two **** likely value. But some software don’t produce this minus two **** value. And the reason we are to multiply these is because the statistical **** says that you can’t just subtract this. If you are subtracting it has to be the minus two **** of the value and when you do that, then it follows the **** square. And what is the difference in number of parameters between this model and this model? It’s one parameters. Right? We have one additional variable in this model that is not in this model. So that’s why we say in **** square with one degree of freedom. So in this case, by adding the pre-test score with high score with one degree of freedom, now what you are basically going to do is to calculate this then we are going to go back to statistics table and then see the **** value for **** square with one degree of freedom. That value is 3.84.

Now, this number is significantly bigger than 3.84 and basically you are saying that, yeah, compared to a model that does not have a pre-test score, the model that has a first test score is a significant improvement. And that’s’ what we are doing. Now, instead of

just looking at this, let's add another variable called the classroom curriculum. Right? We are adding one other variable. And this is the thing that we are getting.

Now what is the meaning of this one is? For one unit increasing "X" and adjusting for the classroom, adjusting for all other variables in the model, this is the increase in the first test score. And what is the interpretation of this? It basically means that adjusting for all other variables in this case pre-test score, this is the change in the **** score for one unit change in the classroom curriculum. So basically compared to people for class curriculum, there are just two levels. Right? They've got the classroom curriculum they did not get the classroom curriculum. Right? Intervention was given or not given. So when I say ****, compared to people who did not get any kind of classroom curriculum, the people who got classroom curriculum their first test score about the effect of tobacco use increased right almost half a point and now we want to see this – is this increase in half a point is it statistically significant. So how do we test it? Again, we do the same thing.

Now we are comparing this model. This is the minus 2 ****, this is the minus 2 **** for this one. So we multiply this by minus 2, we multiply this by minus 2, do our subtraction and we get a value of 30.2. And in this case, yes, compared to a model that just includes pre-test score. If we include the classroom curriculum, that is again, a significant improvement. Compared to a model that just only takes into account the pre-test score, this one makes a significant difference. So whenever we are running this analysis, this is still understood model, but we are including the higher levels of nesting, which is the school. In this case we saw that the pretest score made a significant difference and the classroom curriculum also made a significant difference.

Question? Yes.

Question: ****. If you were **** say that the **** for the pre-score is under ****, would it be the addition of the .3 plus the .4 – if you wanted to see an increase in the .4 ****

Answer: No. When you want to see the increase in both of those ****, basically in the post-test score, whenever you have the **** of all of those two things… oh, the question was, whenever we want to see if there is – what is the total increase. Right?

If you put this value for the individual in the model, in that case you are going to be getting basically. This is all the average; this is not going to be true for each individual. On average, this is going to be that change that we are seeing and this is a plus or randomized model. We are going to run something called the empirical beta estimate. That is going to give us the change and further, each individual also.

Here are on average, this is the change in the first test score when we add the pre-test score, and then this is the increase in the post-test score when we add the classroom curriculum after adjusting for whatever was the pre-test score. But those are average because this is on average for the population that's what we are talking about.

And then… now we see that this cluster level variance it's basically decreasing over here. Right? So what happens is, whenever we are adding these variables, whatever was basically an error in that whatever was not being explained before it's being explained now by the pre-test score. It's being explained now by the classroom curriculum. Right?

So the various **** in the data has now, what happened? Decreased. Right? Because we are at least now, that is the key thing about doing all of this study because every time we are doing a study, we emphasize that we have to be very careful about what really it was we are studying. We want to use variables that explain a lot of variability in the data. And what happens is, when you add these variables in the data, what happens is, it decreases your variance in the data and that's why it's explaining a lot of variability, that's why your cluster level variance is going down and now whenever you are adding variables here to the student level, see in this pre-test score it was at the student level and this class curriculum and it was at the higher level. Right? So when that happens, sometimes it's not just going to be decreasing in one of them, it is going to be decreasing the variable **** in both of them or just one. Right?

So when that happens, it is going to give you a different view of the ICC based upon what you are getting without the variable or what you are getting after the variables. Basically, these ICC is now – this number is now decreasing over here. Why is it happening? Because by putting this variables in our model, we are basically explaining the variable ****, we have taken care of the various **** In the data, now people are becoming more and more similar to each other. Right? And in that case the difference is being taken care of. The difference between the subject is being taken care of. Once the difference between the subjects is being taken care of, **** we are not being able to explain in just a random error, by putting all of these variables in the model we are adding. We are explaining more variability in our data. So that's what – that's why, when you go, we are doing this kind of analysis the importance of having all of this variables in our model is, we want to choose those variables that is going to explain the variability in the data. And we saw that by putting this variable where it explained the cluster variance has decreased, the error variance has decreased and we are explaining the variability in the data, thus reducing the total variability in our data.

Question: I have a question.

Answer: Yes.

Question: What happens how do we interpret when the variance actually increases?

Answer: Sometimes what happens when the variances increase? In this case, usually whenever in the statistical model, whenever you are adding a variable if the variance won't increase, it will give you constant or it will decrease by adding those variability ****. The total variance in your data should not increase. It should always stay constant or decrease and that's how we interpret this "I^2" thing. Right? Whenever we are talking about "I^2" in the Igversion model, that's what we utilize if a model is good or not. So in

that case, once you add a variable, if the **** variable is explaining something in your model, then the"I^2" is going to increase. If it's not adding anything to your model, it is going to stay constant, so that's what is going to happen.

Here in this case, the total variability in our data, one of them increase, the other one should decrease of so that the total variance in the data it's always going to stay constant.

Yes.

Question: So these models, do that consider TV interventions and –

Answer: No. In this case we are just talking about the classroom curriculum. Oh, the person was – are you talking about some other intervention. Right? In this case, this is just for the intervention on the whole analysis of the data. I'm going to show it in a few moments. This one… oh yes.

Question: I'm just wondering about the database structure for this since you have school and then classrooms and then students. Do you typically leave, leave the rows empty underneath or do you repeat the data over and over again, school **** **** cross and then the students?

Answer: Okay. The question was, how do we put the data. Right? Usually the way we set up the data is the school is one, right? So if you have 10 subjects from that one **** classroom is, one, one, one for that one, then we have going from one to 10. Within each classroom there are going to be 20 subjects ****. One, one, one, one, one two, one, one, three, like that. So that the whole subject is repeated like that and that's called the long form of the data if you use SBSS, it's says the wide format and the long format. Long format of the data is when you are listing all the data like that in an elongated format. And when sometimes we call it the wide format, the wide format of the data is if the observances are put next to each other. That is the multi-variant form. And in the uni-variant form for bringing all of these analysis that we need to do in SAS, or in the **** we need to have this data in the long format.

So basically, the school come first, highest level of nesting. Right? And then within the school is the classroom. So, school one, classroom one. Within the classroom whatever the number of observance, we list all of them. And then we list all of it. So, one the higher levels of nesting, it gets repeated, multiple times depending upon how much of that data is there.

And that is one of the biggest, I should say, **** while doing this kind of analysis is whenever we have to do this, if somebody sends me their ****, next year you've got the data already in because I don't want to go through this thing and whenever you are designing your study, whenever you are designing your database, that maybe the best thing to do because if you start with the long format for the data at the beginning of the study, it maybe easier for you to do the analysis at the end if you start with a wide format,

you may have to convert it to the long format and that's a very painstaking **** process. So you don't want to go through that.

Now these things, we are not **** the complete results, we are… the one **** that we are showing this is over here to show you how to do this like **** how to see the model building procedure, but we'll see that in the future… a couple of slides later how this thing is going to change depending upon the ****.

Now instead of students and classroom, we do students in school. So basically the classroom variable has been ignored. So when we do this analysis again, the intercept is very similar, the pre-test score, all of them are similar to each other. And what are the school level to a classroom level or the school level. Adding this pre-test score or the curriculum, all of them it make a significant difference, but what you see there is the high score value has decreased here compared to… decreased here compared to this place. Right. So basically at this level whenever we are going to take into account the curriculum level, the things are becoming a little bit less significant but still that "I^2" with one degree of freedom it's still significant. So whenever you are doing your analysis, what you should notice for this longitudinal ratio test is the "I^2" thing. It's the square quantity.

So it's always going to be a positive quantity. So anybody who is subtracting to get this number, always you should always be getting the positive number. If you get a negative number it means it not going to be "I^2" so you have made an error in your calculations. So when subtracting this, make sure that you are… even if you get a negative umber it means that the absolute value and that should be the positive value and you compared it with "I^2".

Now suppose if I'm comparing this final model with this model. What is going to be the degrees of freedom for the "I^2" here?

Question: Two.

Answer: Two, right.

Because we are adding two more variable over here compared to the model that does not have anything, so whenever you do your "I^2" test, you count how many original data variable you are adding in our model compared to the previous model. And then we call this model as nested model because this model is inside this model. This model is inside this model. And we do like you to **** for models that are only nested in each other. If the smaller model inside the bigger model, then we can do a **** ratio test and sometimes when anybody are doing your analysis, this models won't be nested, so in that case, how do you ****? There is something that the software will produce called the EIC and the BIC, which is the **** Information Criteria and the **** Information Criteria. You don't have to worry about what those are; it will basically be in the software. It's EIC, BIC. And in that case which ever model has the lower EIC or the BIC that's what you use for your analysis.

So in your analysis, if the models are nested, use the **** ratio test. If the models are not nested, use the EIC and the BIC criteria. And I will show those things, what is EIC, BIC how to get those things out from SAS whenever we run the SAS analysis.

Now, going from that model which basically only had the first score, we are not going to ignore the pre-score, but we are going to be adding the all the intervention things for the different levels. Right? So our initial level is a function of the overall level if that due to the CC? If that due to TV and that interacts them between the CC and TV **** basically, what the group who got both of those intervention. And again, our random intercept. Now if cluster effect is completely explained by condescend then this is zero and the variance is zero. Whatever is the condescend if the difference between the cluster is explained by this variables, if these variables are accounting for the variability in the clusters, then we won't need the clusters. Right? So basically this becomes a – the variance of this thing becomes zero. All the estimate of this thing is equal to zero.

Model is the same as ordinarily the **** model. Basically the variables explain everything in the cluster. We don't need the cluster effect; we don't have to say that there is a difference between the clusters to account for that cluster. If that happens then we can just drop out this thing and whenever we chop out this thing, whatever we are running is basically the, the simple regression model where we are saying those are not correlated in the sense.

Now if an **** in for out clusters, then the model is the same as the regressive model and for **** that is the take. If all the clusters are equal means in that case you can treat it as a regressive model. We are just completing the cluster level model. Cluster level means and not have to worry about this doing the multi-level basic thing in this case; we have added a random intercept for the cluster. What is the meaning of this random intercept, it's solves two ****. It basically says that we are taken care of the correlation between the observances by putting this **** and we are trying to… we want to be able to generalize our results. So by putting a random cluster if it for schools or a random cluster if it for classroom, we want to make our results generalizable for the whole group.

Now if you look at this analysis for this ordinary ****, the same analysis we did it as an Igversion analysis and this analysis is for where we are taking into account the cluster effect. In this case, anything that has a value to that isn't void, that is significant at .01 level and some of these other values are significant that .05 level. So whenever we are putting these things you see that all the three models are basically giving the same value for the intercept.

For the classroom curriculum, we see that it's significant for all of them right here in this case, this is a classroom curriculum, it is significant. This is significant, this is significant. And now if you go and look at the residue variance and the class level variance, the residue variance, the more important thing that we want to look at is the class variance. Right? If you look at the class variance that is significant. Right? So, whenever we have a significant cluster if that adding that random effect makes a

difference, you have to take into account that, that means during these ordinary analysis, this may not be really significant.

Now one of the things that you notice over her if you look at this parameter estimates for Television and interaction, between these models. But putting the random effect, what you see here is, this is a negative value. Right? And here is a positive and here is a different positive value. The sign has actually changed between this model and this model for the parameter estimate. You see that same thing over here also.

See, in this case, this model sees that the interact is **** this is significant. Anything that is in italics is significant here. So means… that means the CC by interaction is significant when you look at the student level analysis using that ordinary Igversion model when you don't have that random effect. Once we put that random effect, we see that it is still negative, but it's not statistically significant. And so, between the model where we are using the ordinary Igversion model and the multi-level model. By putting the random effect, which is significant we are getting the result that is in a sense completely different from the results that we would be getting when we are making our analysis just based upon that simple Igversion model.

So when the random effect for this thing is significant, in our case its more meaningful to have a random effect over here than during the analysis. Now if this random effect was not significant, we did not need a random effect, then we would could just do a cross-level analysis or a student level analysis, whatever was your interest, but in this case the data in this **** and we put a random intercept for this things, that means random intercept for the class means we need that class level variances statistically significant. And the reason we are using this analysis is in this case you see that by putting in the random effect we are getting a completely different result that we **** if we assume the data were independent of one another.

Now going to the first model that we did, it was just a… now this is the final model that the **** are used while they were publishing this paper for their analysis. So what was the model, the post-score is a function of the pre-score and then we are taking into account all the things that we are talking about for the intervention right. So what is the interpretation of this **** here?

Which group would we be talking about with this **** over here? Average for the group that did not get any kind of intervention, right. The control group. Basically we are putting zeros over here, right. In our model CC zero means the person did not get any kind of classroom curriculum. TV is zero means they did not get any kind of TV intervention also. Once we put either one of these things or zero this thing is going to be zero for our model. So basically the **** is giving us, what? The **** if we put a random effect also with the zero means we are basically saying that this is the value… initial value for all the people who did not get any kind of intervention and for whom the pre-score was also equal to zero. Right? So that would be ****. And sometimes it may not make any sense, we may get a different value for pre-score, but that is the meaning;

the mathematical meaning of this model is. This **** is going to give us whatever was the value from the group that did not get any intervention.

Now suppose I want to just do it for the people who got the classroom curriculum? How am I going to be changing it means? I am going to put a one for this one right? Classroom curriculum means they got it, means the value is they did not get the classroom curriculum was zero. If they got the classroom curriculum it was one. For TV I am going to say no TV, yes TV, so if I put No TV is equal to zero over here, all of those are going to go away, so what does this Beta2 represent? This Beta2 represents that additional effect that initial difference that you see between that non intervention group and the curriculum group. Right? This is the effect.

So basically, you can interpret all of these things similarly. Now again, for the pre-test whatever is the **** how is the initial score going to affect your score at the **** study. And again in this model we have a random effect which you are denoting by Epsilon, zero I and these are the interpretations for this one.

Any questions still? No?

Question: I just want to make sure they get a break. So,

Answer: Okay. Let's start with, what is the time.

Question: ****

Answer: Okay. So.