# Introduction to Mediation and Moderation

## Video 2 transcript

<https://www.ncrm.ac.uk/resources/online/all/?id=20839>

Oliver Perra: Hello. This is Dr Oliver Perra and this is the second part of my introduction to mediation and moderation. I will firstly summarise some of the key concepts from the first presentation. I have said that mediation answers questions about how predictor influences outcomes, and I have introduced a simple ordinary least square based mediation model where the total effect from X to Y from a predictor X to an outcome Y is partitioned into different effects, a direct effect C, which represents a direct effect of predictor X on outcome Y and an indirect effect represented by the multiplication of paths A and B, which represent how effect of X is transmitted to Y through M. The key inferential test concerned the indirect effect Ae, and this requires resampling methods, for example, bootstrapping or Monte Carlo to estimate the confidence intervals and the distribution of these indirect effects.

[0:01:28]

 In this presentation I will talk firstly about confounding. I will then present a case where the analysis and the mediation model will involve multiple mediators, and I will talk about longitudinal mediation models. It is important to always keep in mind that mediation models are causal models. They make assumptions about causal mechanisms and causal links. Therefore it’s important to think carefully about the causal mechanisms we assume in a model and how we can control for the threats to the validity of the arguments we are putting forward with our models. If the predictor X in a model is an experimental variable, for example treatment that participants receive if they are randomly selected to receive it, as long as participants will receive the treatment and also are otherwise treated in the same way during the study, we can be confident that the treatment is the only plausible cause for differences in the outcomes between the two groups. But often we cannot run experimental studies, and furthermore, while the predictor in an experimental study can be randomly allocated, the mediators cannot. And the mediators are assumed to causally influence the outcome, so the association between the mediators and the outcome may be due to a third variable that influences both.

 I will work again with the example I made in the previous presentation, and this is an example of a treatment that is applied on our soil in order to promote successful growth of our plants. The treatment may work partly because it effectively eliminates fungi and other organisms that hinder plants’ growth. However, there are some possible confounders that may explain this causal link. For example, the pH or acidity in our soil, in our terrain, may explain this causal link because the pH can negatively influence fungi and, on the other hand, the same level of pH can be optimal for plants. This is a fictional example, obviously, but some confounders like that are possible. So, in this case, the indirect effect of treatment through fungi reduction will be biased and confounded. And we would argue that the mechanisms of action of the treatment involve fungi, but this would not be true if some other uncontrolled mechanism so pH can explain the link between the mediator and outcome.

 I take this chance to add an important clarification. We need to be clear about what is being confounded in this example, and here the pH confounder is a confounder of the indirect effect from the treatment to the outcome. In other words, if you were interested in better understanding the mechanism of action of the treatment then you would want to make sure you control for counterarguments that may indicate that the treatment is not actually working by acting on fungi or other mediators, but there may be some other mechanism at play. However, if my main interest was testing if the treatment works and promotes plant growth, whether the mechanism of action is through fungi or involve pH does not really interest me, and all I’m asking is whether the treatment is making the plants grow or not. If the treatment was tested using an experimental procedure on a substantial number of samples, I can be confident it works if I notice that there is better growth after the treatment. The point is that pH, in this example, does not confound the total effect from the treatment to the outcome. If my interest is in estimating the total effect of whether treatment works, I don’t need to adjust for other confounders like these. However, if your interest is understanding the mechanism of action of the treatment then pH is an important source of confounding of the supposed, I professionally call indirect mechanism of (inaudible 0:06:26) reduction of fungi.

 And I will highlight an important resource for reflecting about causal models when you’re building and thinking about the controls you need in your models. This resource is based on directed acyclic graphs, and here I put the link to the webpage daggity.net, which is also available in our package. And directed acyclic graphs are ways to represent directions of causal inferences between variables, and the graphs are called directed acyclic because the edges in the graphs represent a direction of causality between variables and those do not flow back onto themselves, so they are directed and acyclic. These graphs are useful in representing conditions in which we can correctly estimate a causal effect of a predictor on an outcome, and we can add other variables that may be considered as confounders and variables that we can observe, but also there may be confounders that we cannot observe. Here, I represented the example I made just in the previous slide and in this software, in this webpage, I can specify that I want to estimate the total causal effect from the treatment to the outcome growth. And you can see here that if I select that I want to estimate the total causal effect, I am biasing paths that I need to control for. So, if these were all the variables that were important in this particular study then I wouldn’t have to control for confounders if I were just interested in the total effect of the treatment. However, if I want to estimate the direct effect of the treatment or, conversely, the indirect effect, I will have to control for fungus and pH level. The point is that the direct cyclic graphs can help you describe the supposed causal pathway in the phenomenon you are studying and they represent them in a qualitative way, but I can also help you check the type of adjustments and tests that you need to be able to run to identify and estimate correctly the causal hypothesis, so I think they are very useful tools that help you apply your thinking and your hypothesis and also specify what type of controls you need to apply in your models.

[0:09:39]

 So, back to an example of confounding in mediation. In the example in the previous presentation I tested a model whereby the effect of family income on grade 12 math scores was supposed to be mediated by grade 8 reading scores, and again I’ve put some scripts in the material with this module that you can follow so you can run the examples on R. But we can think of many factors that might act as confounders in this pathway. For example, maternal education attainment may influence the child’s course in grade 8 and grade 12, so it might have an impact that confounds the mediating pathway through reading in grade 8. Another variable that may bias this supposed mechanism through grade 8 score is gender. Girls may be more academically driven and do better in grade 8 and 12. So, using the process macro in R, it’s easy to control for similar confounders. All you need to do is add an option C or V for covariates, where you can list a set of covariates. In this case I added male and maternal education as covariates in the model, as you can see in the grey area. Adding controls for these confounders, the indirect estimated effect of high income on math 12 scores is 2.84, whereas it was about 2.82 without those confounders. And note that in the process macro, the total effect of high income on math scores is estimating assuming the covariates are held constant or it is estimated Y controlling for the covariates. The resources provided with this module, you can find more exercises where you can test and apply the process macro in R to estimate different mediation models, so please look at the resources provided.

 Another common scenario in mediation is when parallel mediators are being supposed. In this scenario, the predictor influences the outcome through two mediators or more, and these mediators are supposed not to causally influence each other. And this does not mean that they are independent each from the other; the mediators can be correlated with each other, for example. And this scenario may be particularly useful in testing the different size of indirect effects across different putative mechanisms. For example, here the effect of family high income on grade 12 math scores is supposed to be mediated by grade 8 reading scores and by grade 8 math scores. However, for the sake of example here, I’m not supposing that there is a direct causal link between reading and maths, but this is just for the sake of example. Taking this scenario and using ordinary least square estimation, the values of these mediators are estimated as linear function of the predictor, so the outcome scores are estimated as linear function of the predictor high income, and the sum of coefficients for the different parallel mediators. So, if we have two predictors, as in this case, the grade 12 math scores are estimated as linear function of a direct effect of high income, the C pathway, and the sum of the coefficients, the Bs that represent the effect of the putative mediators on the outcome. Thus the b1 coefficient, for example, represents the changes in grade 12 math scores for one unit increase in grade 8 reading scores for children with the same family income and keeping the grade 8 math scores, the other mediator, constant, so controlling for the other mediator, which means that there are two specific indirect effects in this model, one through grade 8 reading scores, which is given by the product of the paths a1 and b1, and the other specific indirect effect is through grade 8 math scores, which is estimated by the product of a2 and b2. And the total indirect effect is basically the sum of these indirect effects.

[0:15:07]

 So, in the shaded area here I’ve put a syntax that you can use to estimate a similar model. The coefficients estimated through this macro are in the graph here, and the output of the macro can also be used to estimate the indirect effects. So, here we have two indirect effects, one through reading 8 scores and one through math scores in grade 8, and we can see that the indirect effect through reading scores in grade 8 is estimated to be 0.72, whereas the indirect effect through grade 8 math scores is estimated as being 3.20. In the syntax I have also asked the package to estimate contrast, so contrast equals 1 is basically asking a test of difference of the indirect effect, so a test, a new hypothesis that the difference between the two indirect effects is equal to zero. And the results show that we can reject hypothesis that those two indirect effects are equal to zero because the confidence intervals of this test do not cross zero, so we can reject the hypothesis that those two indirect effects are equal, the difference between those indirect effects is equal to zero.

 Using the process macro, it’s also possible to build models that include pathways of inference from one mediator to another, so serial multiple mediators can be included. And here in this example I assume that reading scores in grade 8 casually influences math scores in the same year. Once again, please note that I’m making these assumptions to illustrate how to estimate a similar model, but from a substantive point of view, a similar assumption of a causal link from reading to math scores in the same years is difficult to justify. But if you wanted to test C or multiple mediators then you’ll have to indicate that you want model type 6, so you write model 6 in the syntax of the process macro because this option invokes models with serial multiple mediators. But also note that the order of the mediators after the M option, it’s important by writing read 8 before math 8 I’m instructing the software to consider reading 8 as a cause for math 8, so the order of the mediators after the option M, so specifies the order of the influence in the serial multiple mediators model.

 I will talk about mediation models for longitudinal data, and in one of the references provided with these resources, there is much emphasis on the fact that mediation models applied on cross-sectional studies can only provide accurate estimates of mediation effects in very restrictive conditions. And when these conditions are not met, which is cross-sectional studies will provide biased estimates of mediation effects, and the results can be very misleading, so this makes sense because, as I emphasised on different occasions, mediation models are causal models and they assume some processes of influence and mechanism of change that take time, that necessarily take time to unfold. If, for example, we apply a treatment to our soil to promote plant growth, we will need some time to observe the effect of the treatment on the supposed mediator, the reduction of fungi, and some time to observe the consequential effect that the reduction of fungi has on plant growth. Truly mediation models entail longitudinal data in most cases, so it’s important to consider that and be very careful in making causal assumptions without data and designs that justify those assumptions.

 And with longitudinal data we can also consider more complex situations. For example, here predictor X measured time zero at the start of the study can affect a mediator that we observe at time 1, and the mediator at time 1 can affect the outcome that we observe at time 2. But the previous level of the mediator, the level of the mediator at time zero at the start of the study can also affect the mediator at time 1, and similarly the previous level of the variable Y at time zero and time 1 can have an effect that we observe at the end of the study at time 2. So, these are autoregressive effects that reflect the stability of individual differences across time. And if we can control for these effects, we can also have estimates of the direct and indirect effects that are more reliable, so these more complex models can be estimated using structural equation models, for example, which provides a more flexible approach to mediation analysis, one where different types of variables can also be included.

 And indeed, panel data can be used to test more complex models like these. In this case the three variables are collected at three timepoints, and note that unless the predictor X is a variable that we can control experimentally, then we also need to model the correlation between the predictor and the other variables at time zero at the start of the study, and these correlations are represented by the grey arrows at different timepoints. We can then assess the autoregressive effects of the three processes here, indicated by the dotted lines. And the effect of X from X to the mediator Ma are assessed on two occasions, as are the effects from the mediator to the outcome. So, we have different estimates of the pathways from the predictor to the mediator and from the mediator to the outcome and while, at the same time, we control for stability of individual differences in the process. So, using path analysis and structural equation analysis, we can also use this model to test the stationarity of the causal processes, for example, if the magnitude of effects from predictor aX to mediator M and from the mediator M to the outcome Y change across time, and we can also test other hypotheses about the differences, the changes in variances and covariances across time here. Those variances and covariances are represented by the grey circles and arrows. So, we can test more complex models that can also give us more reliable estimates of and also more complex estimates, more information about the pathways from supposed cause to mediators and outcomes.

[0:23:57]

 Here, I put a reference that used this approach. In this study the researchers showed anxiety sensitivity, which is the tendency to interpret unpleasant physiological sensation as dangerous, prospectively predicting alcohol problems but not alcohol consumption across high school. Interestingly, anxiety mediated the anxiety sensitivity effect on alcohol problems across high school, so you can look at this reference to see an example of how this approach can be successfully used to provide evidence on mediation mechanisms and how some predictors can affect outcomes through other constructs.

 The use of structural equation models also allows us to integrate other types of models into mediation analysis. For example, if the moderator is observed over repeated periods of time, we can use latent growth curve models to estimate the interceptor slope of these outcomes, so the initial status and the rate of change of the moderator. And these parameters that represent change in the moderator over time can then be included in the mediation model and test if changes in the mediating variable are responsible or transmit some of the effect of the independent variable on the outcome. And I refer also to resources I have created for the National Centre for Research Methods on latent growth curve models.

 Finally, it is also possible to use mediation models and apply them in situations where the predictor and the mediators are at different levels compared to the outcome. For example, in this scenario we have children who are clustered within classes in school. The intervention may be delivered at the level of classes, and the moderating mechanism that we hypothesise can transmit part of the effects of the intervention may be also the class level. For example, the intervention may work by improving teacher-pupil communication so the intervention is a moderator that is also at class level and this moderator in turn influences individual children’s results. There are packages that can run these type of analyses. However, structural equation models that also include and integrate generalised linear models provide a more flexible approach, one that allows to consider observed as well as latent variables in the mediation model and can include different types of variables, for example, binary ordered categorical counts(?).

 So, thank you for your attention, and please look at the exercises provided with this module, which will guide you in building different mediation models using (inaudible 0:27:39). Thank you again. Bye.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK