Preliminary Analyses for Social Phobia Example

In this document, I describe the preliminary analyses I routinely use for RETs with continuous mediators and outcomes. After providing a refresher of the numerical example, I consider the analysis of response distributions, evaluation of treatment imbalance, matters of non-linearity, outlier/leverage analyses, choice of covariates, and omitted moderation. Technically, SEM based maximum likelihood estimation

THE NUMERICAL EXAMPLE

The RET explores a two group (treatment versus control) design to reduce social phobia. Social phobia is a mental health condition characterized by intense anxiety about social situations that leads to significant impairments in everyday life. The program targeted three mediators/mechanisms. The first mediator is negative cognitive appraisals. People with social phobia believe they will behave ineptly and unacceptably in social situations and that doing so will lead to loss of status, loss of worth, and rejection. The program sought to reduce such negative appraisals. The second mediator is perceived social skills, i.e., people's perceptions of their ability to manage potential threats in social situations. The program sought to increase confidence in one's social skills. The third mediator is external locus of control in social situations. This refers to beliefs that events during social interactions are controllable only by people other than oneself, leading to a sense of lack of predictability and control. The program sought to decrease such feelings. The control group was a wait list control but received some educational materials about dealing with social phobia during the wait period.

Three interchangeable indicators of the outcome were measured at baseline and again three months after program completion. One measure was a variant of the Social Phobia Inventory (SPIN), a patient self-report of social phobia symptoms. Multiple symptoms are rated on a metric indicating how often they occurred during the past week (0 = never, 1 = very infrequent, 2 = infrequent, 3 = sometimes, 4 = frequent, 5 = very frequent, 6 = always). The second measure was the Social Phobia and Anxiety Inventory (SPAI), a multi-item self-report of symptoms. Individuals rated items on the same metric as SPIN. For both measures, a total score was defined as the average of item scores for an individual across items. The third measure of social phobia was a clinician rating based on an extensive and systematic clinician interview with the patient. The rating was made on a six-point metric with the values 0 = not social phobic, 1 = mild social phobia, not disabling, 2 = moderate social phobia, somewhat disabling, 3 = social phobic, moderately disabling, 4 = quite social phobic, quite disabling, and 5 = extremely social phobic, very disabling. Clinicians could assign decimals to make finer gradations. I treat the measure as interval enough for analytic purposes.

The mediators were measured at baseline and program completion. Each was measured using a multi-item inventory with responses to items on 7 point agree-disagree scales: -3 = strongly disagree, -2 = moderately disagree, -1 = slightly disagree, 0 = neither agree nor disagree, 1 = slightly agree, 2 = moderately agree, 3 = strongly agree. The scores were averaged across items. Higher scores imply greater negative cognitions, perceived social skills, and external locus of control. The N was 333.

To keep matters simple for purposes of pedagogy, I limit the number of covariates in the example. For the mediator-outcome portion of the model, I focus on two confounders that prior research suggests might artificially inflate the association between each mediator and the outcome. The first confounder is biological sex. Research indicates there are sex differences in social phobia (females suffer more from social phobia than males) as well as sex differences in each mediator. The second confounder, measured at baseline, is the extent to which patients grew up with parents who were hypercritical of them. Prior research suggests that such a family history influences each of the mediators and social phobia, again taking on the role of a confounder. This covariate was measured on a multi-item self-report where each item was rated on a -3 to +3 disagree-agree metric. Items were averaged. Higher scores indicate a greater family history of hypercriticism.

The RET model appears in Figure 1, absent covariates to avoid clutter. In the figure, the number 1 after a variable name indicates a baseline assessment, 2 indicates an immediate posttest assessment, and 3 indicates an assessment 3 months after treatment completion. I notate the path coefficients with numbers after the letter *p*. I use the letter *d* to signify disturbance terms and *e* to signify measurement errors. In terms of the covariates, I use biological sex and parental hypercriticism for each endogenous variable in the model. I also control for the baseline variable of the modeled endogenous variable.

As discussed in Chapters 3 and 7, when working with latent variables, we need to assign a metric to them. I used the clinician rating as the reference variable and pass its metric to the latent variable using the methods discussed in Chapter 7. The rating ranges from 0 to 6 with clearly demarcated and intuitive reference points; 0 = not social phobic, 1 = mild social phobia, not disabling, 2 = moderate social phobia, somewhat disabling, 3 = social phobic, moderately disabling, 4 = quite social phobic, quite disabling, and 5 = extremely social phobic, very disabling. The metric of latent social phobia can be thought of in these terms, adjusted for measurement error per the error theory within the model.

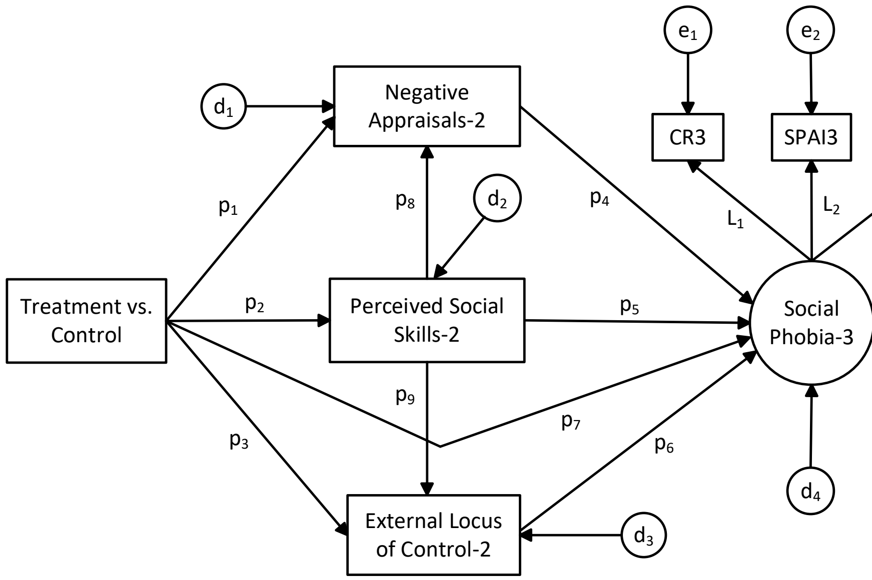

FIGURE 1. Social phobia example

THE MODEL EQUATIONS

It will be helpful to translate the influence diagram into the implied linear equations but also incorporate the covariates into them. I use *p* notation for the path coefficients and *b* notation for coefficients associated with covariates. I invoke the heuristic that expresses each endogenous variable to be a linear function of all constructs with arrows pointing directly to the endogenous variable. Here are the equations using sample notation (I use short labels for the variable concepts to save space; I use somewhat different labels later for the *measures* of the concepts. The codes are T = treatment condition, PSS = perceived social skills, NCA = negative cognitive appraisals, ELC = external locus of control, LSP = latent social phobia, BS = biological sex, PH = parental hypercriticism):

$$NCA2 = a_1 + p_1\,T + p_8\,PSS2 + b_1\,BS1 + b_2\,PH1 + b_3\,NCA1 + d_1 \quad [1]$$

$$PSS2 = a_2 + p_2\,T + b_4\,BS1 + b_5\,PH1 + b_6\,PSS1 + d_2 \quad [2]$$

$$ELC2 = a_3 + p_3\,T + p_9\,PSS2 + b_7\,BS1 + b_8\,PH1 + b_9\,ELC1 + d_3 \quad [3]$$

$$LSP3 = a_4 + p_7\,T + p_4\,NCA2 + p_5\,PSS2 + p_6\,ELC2 + b_{10}\,BS1 + b_{11}\,PH1 + b_{12}\,LSP1 + d_4 \quad [4]$$

$$CR3 = a_5 + L_1\,LSP3 + e_1 \quad [5]$$

$$SPAI3 = a_6 + L_2\,LSP3 + e_2 \quad [6]$$

$$SPIN3 = a_7 + L_3\, LSP3 + e_3 \quad [7]$$

$$CR1 = a_8 + L_4\, LSP1 + e_4 \quad [8]$$

$$SPAI1 = a_9 + L_5\, LSP1 + e_5 \quad [9]$$

$$SPIN1 = a_{10} + L_6\, LSP1 + e_6 \quad [10]$$

THE PRELIMINARY ANALYSES

It generally is good practice to check key assumptions of one's modeling approach. I use robust estimation algorithms for model estimation, so traditional assumptions of non-normality and variance heterogeneity are of lesser concern. However, some distribution shapes can impact how I choose to model data, such as the presence of sparse data or highly skewed, asymmetric data with non-trivial outliers. For nominal variables, I routinely examine frequency distributions of them to determine if there are base rate issues (the data are congregated in only one or two categories) or sparse data to contend with. This was not a problem for the current example. For continuous or many-valued quantitative variables, I often construct kernel density plots to gain an appreciation for the distribution shape. **Kernel density plots** are smoothed histograms. Rather than plotting frequencies, they plot densities across the different values of the distribution. Unlike histograms, kernal density plots are not affected by the number of bins used to define groupings and often better reflect the shape of the distribution (although density plots also have their limitations; see Wilcox, 2021). My website has a program called *density plot* for creating such plots. Figure 2 presents the density plot for perceived social skills at time 2 (called `PSKILLS2` in the data). The distribution has a bimodal-like shape. This occurs because the distribution is a mixture of two normally distributed variables, the posttest perceived social skills for the control group and the posttest perceived social skills for the treatment group. The latter group's mean has been shifted to the right because of the intervention, hence the distribution shape in Figure 2. The resulting non-normal distribution is not a major concern because what matters most is the distribution of the disturbances when predictors in an equation are held constant and also because of my use of robust estimation. Nevertheless, it is good practice to inspect and take into account, as needed, the variable distributions you work with. Figure 3 shows the density plot for perceived social skills as measured at baseline. The density plot is symmetrical and bell-shaped because it is not subject to the mixture dynamic mentioned above. I examine frequency distributions and plots for all of the variables, but in the interest of space, I only show the one here.

Linearity and Correlation Analyses

My analytic strategy often assumes linear relationships between the continuous or many-valued quantitative mediators, covariates, and outcomes. To check the viability of such assumptions, I create smoother plots between such variables using the program for bivariate smoothers on my website called *bivariate smoothers* (described in Chapter 6). For example, I assume that perceived social skills at the posttest is linearly related to social phobia such that higher levels of perceived social skills are associated with lower levels of social phobia. One wrinkle to evaluating the viability of this assumption in the case is that social phobia is a latent variable. In such cases, I construct separate smoothers for each indicator of the latent variable. In that way, I can check if each indicator is related to perceived social skills in the way I expect, i.e., linearly. Figure 4 shows the plot for the clinician rating (`CR3`), which is the reference indicator, and `PSKILLS2`. I look for a reasonably straight line on the plot.

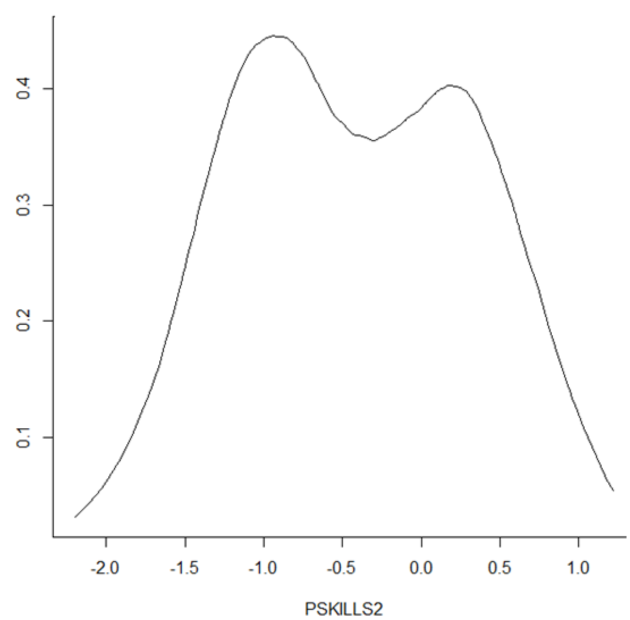

FIGURE 2. Kernel density plot for perceived social skills at posttest

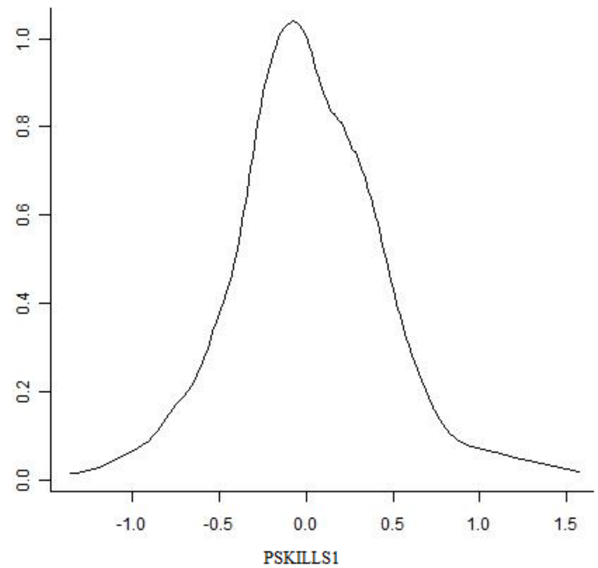

FIGURE 3. Kernel density plot for perceived social skills at baseline

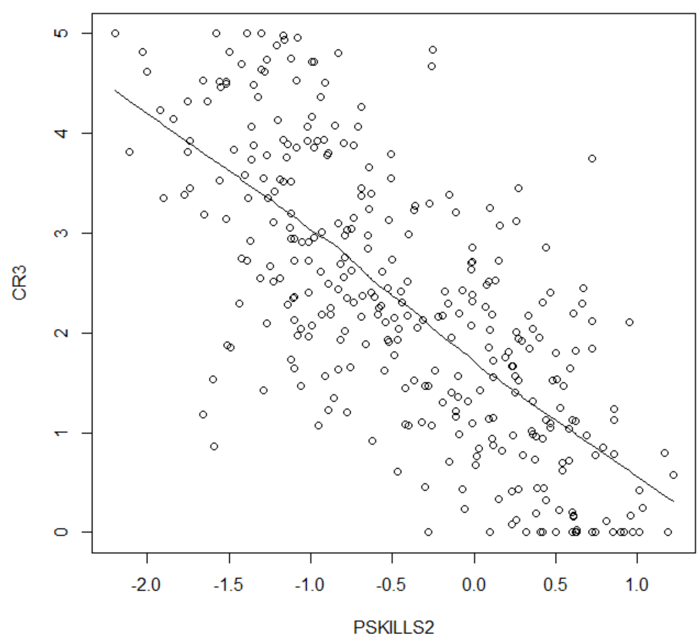

FIGURE 4. Smoother for perceived social skills and clinician ratings

On my website is a program called *scatterplot matrix* that provides a rough picture of variable distributions and smoothers on a single multivariate plot. I show in Figure 5 of a scatterplot matrix using an example with the three posttest mediators, labeled `NEGAPP2` for the posttest negative cognitive appraisals, `PSKILLS2` for the posttest perceived social skills, and `EXTERN2` for the posttest external locus of control, and the clinician rating of social phobia measured at time 3. The diagonal of the matrix has a histogram for each variable, the correlations between the variables are in the upper triangle of the matrix, and the smoother plots are in the lower triangle of the matrix. The plots are not as fine-grained as the those in Figure 2 through 4, but they provide a quick sense of the underlying dynamics. All appears to be in order in terms of linearity vis-a-vis the plots and this was true when I examined other variables in similar plots.

Linearity between a predictor and an outcome often is subject to controlling for other predictors in the target equation. It is possible for a non-linear relationship to turn into a linear relationship when covariates are controlled, just as a linear relationship can turn into a non-linear relationship when covariates are controlled. On my website is a program called *partial residual plots* that creates covariate-controlled smoothers (see Chapter 6 for details).

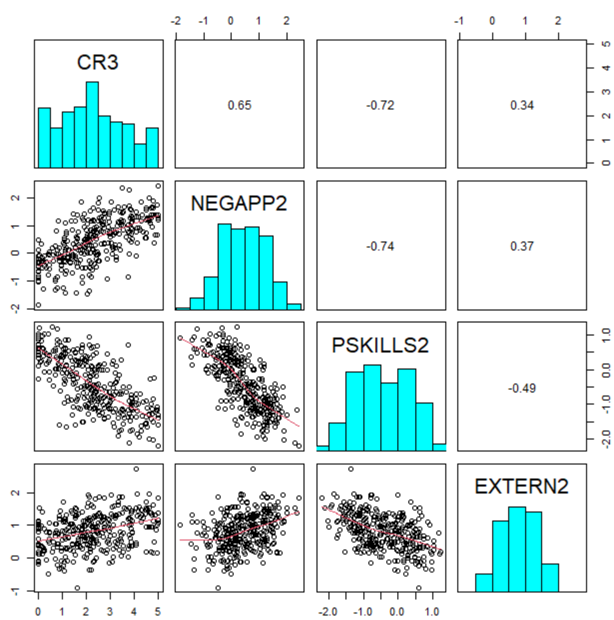

FIGURE 5. Example scatterplot matrix

Figure 6 presents the component plus residual plots for each of the three mediators predicting the clinician rating at time 3 but where all three mediators serve as predictors coupled with the baseline clinician rating, biological sex, and parental hypercriticism, per Equation 4 but where I use CR3 in place of the latent social phobia variable. These plots evaluate linearity in predictor-outcome relationships holding constant the other predictors in the regression equation. This is important to evaluate because seemingly non-linear relationships can become linear when other variables are constant and vice versa. Figure 6 presents the plots for the three mediators in the model. The X axis in Figure 6 represents scores for the target predictor/mediator. The Y axis is the regression coefficient for the target predictor times the person's score on that predictor, called the component value; we add to this value the person's residual score from the full analysis. The result is the **component plus residual value** for an individual.

The residuals in the component plus residual value contain within them the influence of all other independent factors that influence the outcome other than the linear predictors in equation. This includes any operative non-linearities from the target predictor, which are ignored in the primary regression analysis because of its exclusive focus on linearity. To these residuals, we add back the (covariate adjusted) linear contribution of the target predictor using the component portion of the component plus residual term. The resulting score is a mix of the linear and non-linear influence of PA2 on the outcome.

The plot In Figure 6 shows the best fitting line between the component plus residual values and each mediator. It is the dashed lines. The figure also plots a solid line smoother for the data, which captures the combined linear and nonlinear influence of each mediator. If the smoother is functionally linear and overlays the dashed line, then this implies the mediator is linearly related to the outcome. If the lines diverge substantially, this implies non-linearity.

To avoid the trap of overfitting, I do not pay much attention to minor deviations from linearity and I also keep in mind that smoothers are not perfect; sometimes at extreme values they are impacted by data sparseness. The slight flattening of the smoother at the low end of NEGAPP2 and the upward bend at the higher values of EXTERN2 catch my attention, but I would be reluctant to act on them because (a) they are outside the bulk of the data cloud where linearity dominates and (b) they do not make much theoretical sense. I generated the population data to be linear in structure, so I know that the non-linearity I am observing in the plots is due to sampling error. However, in practice, I would not know this. The polynomial analyses I report next are nice complements to inspecting the partial residual plots because they take sampling error into account in a stronger way.

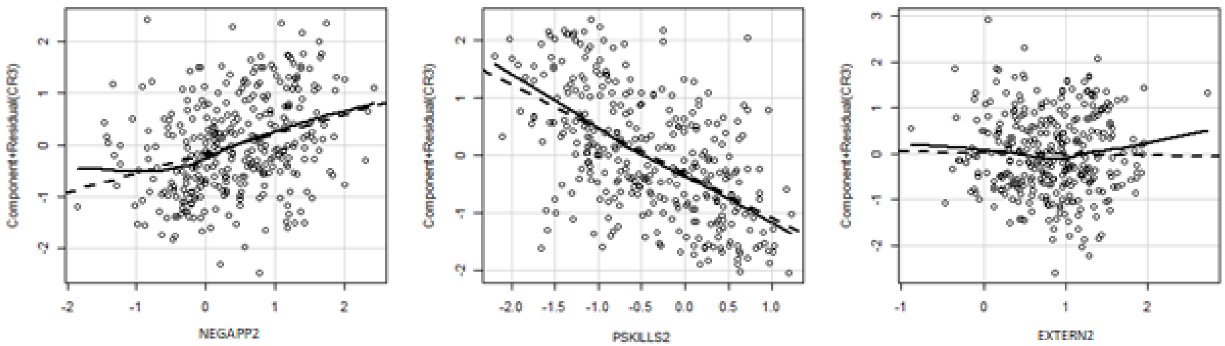

FIGURE 6. Partial residual plots

The second method I use is to evaluate curvilinearity taking into account covariates is to target one quantitative predictor at a time and then add a quadratic term (the predictor squared) to the equation and evaluate the statistical significance of its coefficient. I then add a cubic term (the predictor cubed) to this model and evaluate the statistical significance of its coefficient, followed by a quartic polynomial and then a quintic polynomial. Here is an example of two equations I evaluated for the perceived social skills (PSS2) mediator predicting the clinician rating (CR3) for the case of quadratic and cubic terms:

$$CR3 = a_1 + b_1\ T + b_2\ NCA2 + b_3\ PSS2 + b_4\ PSS2^2 + b_5\ ELC2 + b_6\ BS1 + b_7\ PH1 + b_8\ CR1$$

$$CR3 = a_2 + b_9\ T + b_{10}\ NCA2 + b_{11}\ PSS2 + b_{12}\ PSS2^2 + b_{13}\ PSS2^3 + b_{14}\ ELC2 + b_{15}\ BS1 + b_{16}\ PH1 + b_{17}\ CR1$$

For the first equation, if b_4 is statistically significant, this implicates a single bend curve for PSS2. For the second equation, if b_{13} is statistically significant, this suggests a two-bend curve. Note that I do not focus on the significance patterns of lower order coefficients, only the coefficient for the highest order polynomial in the equation. See Chapter 6 and X for details. On my website, I provide a program called *polynomials* that generates all polynomial terms and evaluates coefficients up to the fifth order. None of the higher order coefficients were statistically significant in the social phobia example. Coupled with the partial plots and smoothers, linear functions seem reasonable.[1]

[1] I often repeat the analyses with the other latent indicators or I use latent non-linear modeling, but the latter is a bit involved.

Outliers and Leverages

Next, I perform checks for outliers and extreme leverages. There have been several suggestions for how to identify influential cases in SEM models (Pek & MacCallum, 2011; Sterba & Pek, 2012) but most are subject to outlier masking, as discussed in Chapter 6. They also are cumbersome. Outlier analysis in FISEM can be pursued at the level of specific model equations or it can focus on the overall multivariate pattern of scores for the full set of variables in the model. It is my experience that the former approach usually is more informative, so I illustrate it here. I use a LISEM mindset for these analyses because the approach I use is not available for FISEM. Although mixing FISEM and LISEM conflates analytic approaches, I think the strategy is better than using the less-than-ideal outlier analysis methods currently available for FISEM. For a discussion of outlier effects on global fit indices, see Yuan and Zhong (2013).

For each equation in my model, I apply the Rousseuw and van Zomeren method discussed in Chapter 6. Again, for the latent variable, I choose the reference indicator (CR3) as my outcome and then replicate the analysis for the other indicators of social phobia. Figure 7 presents the relevant plot predicting CR3 from the predictors in the LSP3 equation but substituting CR1 for LSP1. Potentially problematic cases appear in the upper and lower right quadrants (see Chapter 6). There are none so no corrective actions are needed. If problematic cases are identified, I make a list of all the problematic cases across the equation-by-equation analyses. I then compare the primary modeling results with and without those cases included to see if the results are comparable. If they are, I report the results with the problematic cases included. If not, I might shift to LISEM using outlier resistant, robust regression methods.

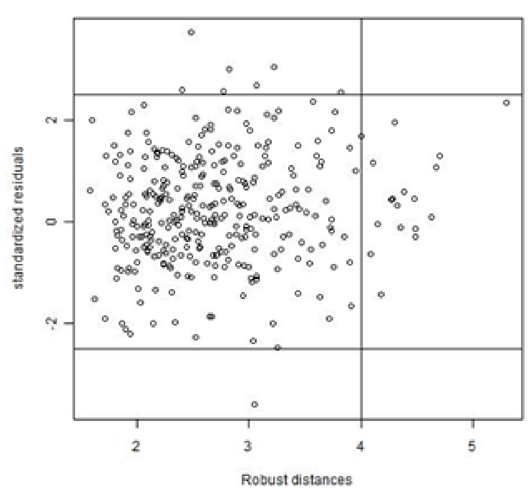

FIGURE 7. Regression outlier/leverage analysis

Covariates

It is not uncommon in RCTs and RETs with continuous outcomes to include the baseline mediator as a covariate for its respective posttest mediator and the baseline outcome as a covariate for the posttest outcome. As discussed in multiple chapters in the main text, this often increases statistical power, adjusts for sample imbalance, and helps control for distal unmeasured confounds. However, mediational analyses for the direct effect of the treatment on the outcome over and above model mediators raise the possibility of collider bias. I discuss colliders in Chapter 2. I like to explore this possibility when evaluating direct effects of the treatment independent of predictors, either as a form of preliminary analysis or as a form of sensitivity analysis. I illustrate my approach using the social phobia example, but I suggest you review the material in Chapter 2 on colliders to assist understanding of this section.

Mayer et al. (2014) present a stochastic theory of mediation to help appreciate collider bias and covariate selection when evaluating the direct effect of a treatment on an outcome holding constant the mediators and the baseline covariates. For the Chapter 11 example, this involves evaluating Equation 4 from above, which I repeat here for convenience:

$$LSP3 = a_4 + p_7\ T + p_4\ NCA2 + p_5\ PSS2 + p_6\ ELC2 + b_{10}\ BS1 + b_{11}\ PH1 + b_{12}\ LSP1 + d_4$$

where all terms are as previously defined. Interest is in estimating p_7. I include the three mediators negative cognitive appraisals, perceived social skills, and external locus of control in the equation as well as the covariates biological sex, parental hypersensitivity and the latent baseline outcome. The choice of these covariates is based on theory and by the general practice of including the baseline measure of the outcome as a covariate to account for distal unmeasured confounders. Note, however, that there is a broader pool of measured covariates I could include in the equation, namely the baseline measures I have for each mediator. If I include, for example, the baseline perceived social skills measure, PSS1, as a covariate note that I now have within the predictor pool two determinants of PSS2, namely T and PSS1. If I also hold constant the mediator PSS2 in the equation analysis, this sets up a collider dynamic that can impact my estimate of p_7 because I am covarying out a collider, PSS2. In theory, T and PSS1 should be zero correlated because I randomly assigned participants to T. As described in Chapter 2, the inclusion of PSS2 in the equation could produce a false partial correlation between T and PSS1 in the context of the equation which, in turn, can bias the estimated effect of T on LSP3. Mayer et al. (2014) discuss in depth the underlying logic and describe methods for exploring if collider bias is consequential for the analysis of the direct effect of T on the outcome.

A strategy I find helpful is to estimate p_7 in the above equation with and without the covariate in question (in this case, PSS1) to determine if its inclusion substantively changes the value of p_7. Here are the results from the original analysis as reported in the main text (with p_7 highlighted in red):

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
LSP3 ON				
LSP1	0.347	0.072	4.835	0.000
NEGAPP2	0.390	0.095	4.100	0.000
PSKILLS2	-0.707	0.099	-7.109	0.000
EXTERN2	-0.002	0.091	-0.017	0.986
TREAT	**-0.488**	**0.136**	**-3.581**	**0.000**
SEX	-0.002	0.088	-0.026	0.979
HYPER	-0.186	0.103	-1.803	0.071

Here are the results with the baseline measure of perceived skills (`PSKILLS1`) added as a covariate:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
LSP3 ON				
LSP1	0.350	0.075	4.697	0.000
NEGAPP2	0.392	0.096	4.089	0.000
PSKILLS2	-0.715	0.103	-6.931	0.000
EXTERN2	-0.001	0.091	-0.016	0.987
TREAT	**-0.475**	**0.148**	**-3.217**	**0.001**
SEX	-0.001	0.089	-0.006	0.995
HYPER	-0.181	0.103	-1.757	0.079
PSKILLS1	0.026	0.112	0.234	0.815

The estimate of p_7 is close to that of the original analysis as are the results for the other predictors in the equation. Whatever collider effects are operative appear to be inconsequential for the path coefficients that are of primary interest. This also was true when I explored the baseline covariates for the other mediators, both one at a time and multivariately. This makes me less concerned with collider bias.

For a fuller discussion of the above tests, see Mayer et al. (2014).

Additional Analyses

For a discussion of additional preliminary analyses you might consider, watch the video on my programs tab for the program titled *Regression Diagnostics*. Also, I generally make it routine practice to determine if two way moderation exists among predictors in each linear equation within my model to ensure I have not omitted important moderated effects that might lead to omitted variable bias. I also check quantile treatment effect plots to determine if the specialized form of moderation for these methods is present. Watch the video for my program called *Quantile Plots* on the programs tab of my website.

REFERENCES

Mayer, A., Thoemmes, F., Rose, N., Steyer, R. & West, S. (2014). Theory and analysis of total, direct, and indirect causal effects. *Multivariate Behavioral Research*, 49, 425-442.

Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research*, 46, 202–228

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods*, 17, 582–599.

Yuan, K. & Zhong, X. (2013). Robustness of fit indices to outliers and leverage observations in structural equation modeling. *Psychological Methods*, 18, 121–136.