11

Mediation Analysis with Continuous Outcomes

There is nothing so stable as change

- BOB DYLAN

INTRODUCTION

In this chapter, I develop mediation analysis in RETs that have continuous mediators and continuous outcomes. I show how to program Mplus for maximum likelihood based full information SEM (FISEM) and Bayesian SEM. I also apply limited information SEM (LISEM) to RET data and briefly discuss RET analysis for Pearl's causal mediation framework. I present an RET influence diagram, derive model equations from it, conduct preliminary analyses, conduct the mediation analyses, and interpret the Mplus output. I assume you are familiar with Chapters 1 through 10. The Chapter is long because it

reproduces syntax and output from the worked examples. The description of maximum likelihood based FISEM analysis is extensive, with descriptions of the other analytic approaches (Bayes FISEM and LISEM) being briefer but readily applicable to extended RET analyses. Each section can be read in separate sittings.

WORKED EXAMPLE

The example in this and other chapters uses simulated data. The example RET invokes a two group (treatment versus control) design to reduce social phobia. Social phobia is a 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 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 not controllable by oneself, leading to a sense of lack of predictability. The program sought to decrease such feelings. The control group received superficial educational materials about social phobia.

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, researchers typically sum scores across items to yield a total score. As discussed in Chapter 2, I prefer to average items of multi-item inventories. By averaging, the total score is tied to the item metric as a reference point. Someone with a total score of 4.5 on the SPIN, for example, tends to rate items in the "frequent" to "very frequent" range of the 0 to 6 metric. Someone with a total score of 0.6 tends to rate items in the "never" to "very infrequent" range. This scoring strategy does not alter significance tests but makes the scale more interpretable. I averaged responses for SPIN and SPAI.

The third measure of social phobia was a clinician rating based on an extensive clinical 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 judgment gradations. I treat the measure as interval enough for analytic purposes. Given the many values, it would be difficult to analyze ordinally.

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 I use. For the mediator-outcome portion of the model, I include 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.

Although I include these two covariates in the analysis, some might argue that I do not need them. To increase statistical power and adjust for sample imbalance in the treatment versus control condition, I use the baseline social phobia as a covariate when analyzing mediator-outcome relationships. In Chapter 2, I described the strategy of controlling for distal confounders by controlling proximal confounders that block the pathways through which the distal confounders affect M or Y. For example, it is likely that both biological sex and parental hypercriticism influence posttest social phobia but only through their effects on baseline social phobia, per paths *a*, *b* and *c* in Figure 11.1. By controlling baseline social phobia, I block these pathways, rendering the two more distal covariates harmless. The reason to include biological sex and parental hypercriticism as covariates is if I believe they impact posttest social phobia over and above their impact on baseline social phobia vis-à-vis the dashed arrows in Figure 11.1. This seems unlikely. I will go ahead and include them to illustrate how to handle covariates in RET analyses. Indeed, I will include the two covariates for all the endogenous structural variables in the model to illustrate the fundamentals of working with covariates. However, in practice, you will give careful thought to covariate inclusion per my discussion in Chapter 2 and below. I also control for the baseline variable of the

modeled endogenous variable. For example, for the posttest negative appraisals, I control the baseline negative appraisals in the spirit of an ANCOVA model.

RET Model

The RET model I evaluate appears in Figure 11.2, absent covariates to avoid clutter. In the figure, the number 1 after a variable name indicates a baseline assessment, 2 is for the immediate posttest, and 3 is for 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.

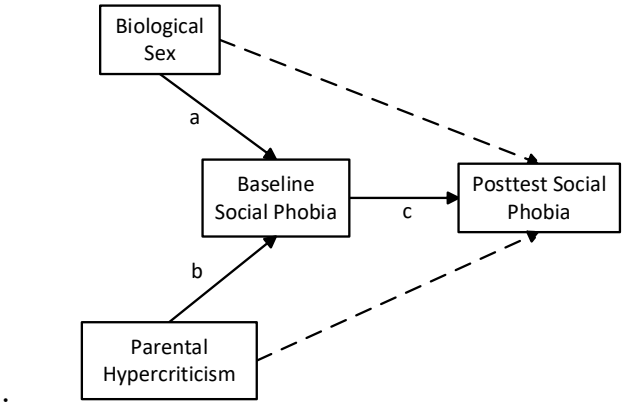

FIGURE 11.1. Controlling baseline outcomes (disturbance terms are omitted for clarity)

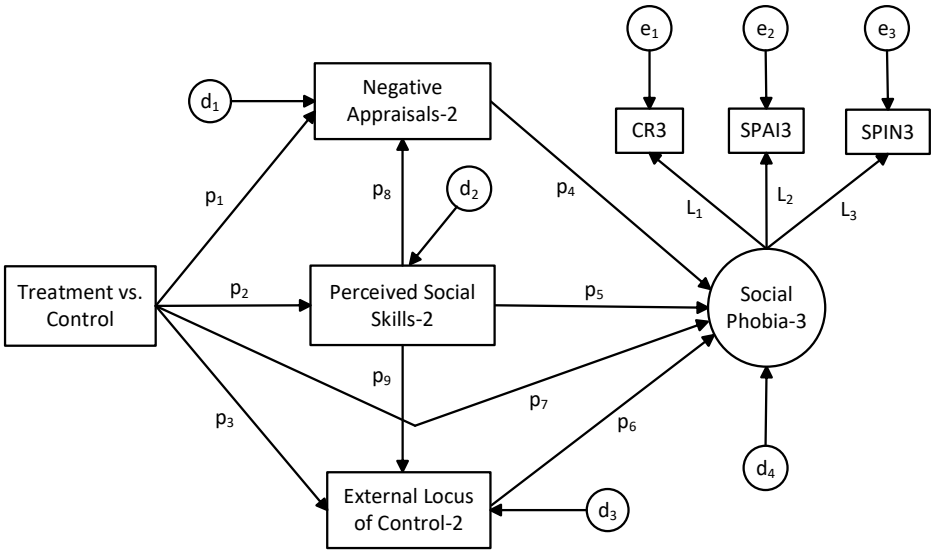

FIGURE 11.2. Social phobia example

Causal Relationships Among Mediators and Correlated Disturbances

Based on past research, the RET model in Figure 11.2 posits causal effects among some of the mediators. Specifically, the model specifies that the better people think their social skills are, the less they think that negative social consequences will happen in social situations (p_8). As well, the better people think their social skills are, the less they will attribute what happens to them in social situations to factors out of their control (p_9).

An important step when formulating an RET model is to think through the need for correlated disturbances. I illustrate why this is important using the influence diagram depicted in Figure 11.3a., which focuses on only a portion of the example RET model. The diagram shows only the causal effect of perceived social skills (PSS2) on negative cognitive appraisals (NCA2) at the posttest. Figure 11.3a implies that perceived social skills and negative cognitive appraisals are correlated for one and only one reason, namely because PSS2 causes NCA2. Indeed, we ultimately will infer the strength of the causal path linking the two variables by examining the magnitude of the association between them. The disturbance terms in this figure represent unmeasured causes of PSS2 and unmeasured causes of NCA2, respectively. Note that the two disturbance terms are assumed to be uncorrelated. If some of the unmeasured causes of PSS also are unmeasured causes of NCA, then the disturbance terms should be positively correlated, per Figure 11.3b. This is because each disturbance term shares one or more of the same (unmeasured) common causes.

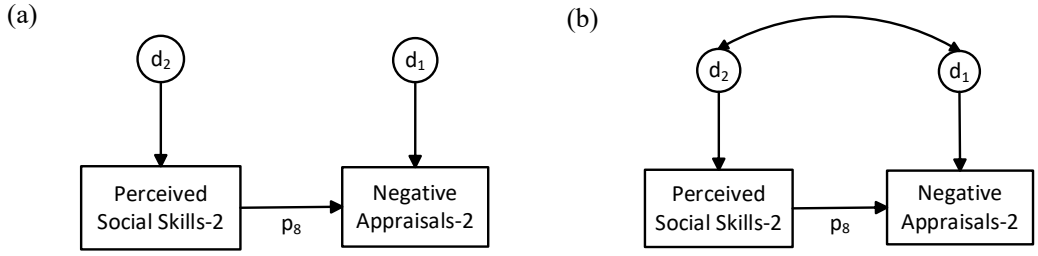

FIGURE 11.3. Causal relationships among mediators

In Figure 11.3b, I have added a correlation between the disturbances on the assumption that there are unmeasured shared common causes in them (e.g., perhaps ethnicity, or SES, or whatever). For this model variant, there are now *two* sources of the correlation between PSS2 and NCA2, (1) the causal effect of PSS2 on NCA2, and (2) one or more unmeasured common causes of PSS2 and NCA2 that reside in each of the two

disturbance terms. If my intent is to infer the strength of the causal effect of PSS2 on NCA2 from the magnitude of the association between the two variables, I need to factor out the contributions of the unmeasured common causes to the correlation because they are a spurious source of association between them. It turns out that specifying the disturbances as correlated in the model will allow me to do so vis-à-vis the mathematical machinery of SEM. This is a strength of SEM. Note that if I fail to include a correlation between the disturbances when such unmeasured common causes exist, the model will be misspecified and I will over-estimate the causal effect of PSS2 on NCA2 by attributing all of the association between the variables to the casual effect of PSS2 on NCA2. Modeling correlated disturbances is important if they do indeed exist.

In the current RET, I explicitly measured and controlled for two plausible common causes of perceived social skills and negative cognitive appraisals, namely biological sex and hypercritical parenting. By measuring and controlling for these two covariates for both PSS and NCA, I essentially lift them out of the disturbance terms and thereby lessen the need to correlate the disturbances. These two variables are no longer unmeasured confounds. By also covarying out the baseline measure of perceived negative cognitive appraisals, I block the pathways of distal confounders that influence NCA2 through NCA1, further reducing the need to correlate the disturbances. One idea behind the inclusion of such covariates is to reduce the correlation between the disturbances to zero, thereby negating the need to include a correlation between the disturbances in the model. I can then proceed with the analysis without correlated disturbances, which greatly simplifies the underlying mathematics. Having said that, I am always wary that there may exist other unmeasured confounds that I have not thought of and measured that can wreak havoc with my inferences. Later, I show you how to conduct sensitivity tests to evaluate this possibility.

Note that for negative cognitive appraisals and external locus of control at the posttest, the correlation between d_1 and d_3 in Figure 11.2 also is presumed to be zero because there is no curved arrow connecting d_1 and d_3. Again, the idea is that all the variables that meaningfully account for the correlation between negative cognitive appraisals and external locus of control are represented in the model so I do not need correlated disturbances. For example, one reason negative cognitive appraisals and external locus of control are correlated is because of the common cause of the treatment condition on them (p_1 and p_3). Another reason is the common cause of perceived social skills (p_8 and p_9). Yet another reason is the common cause of biological sex and parental hypercriticism. A final reason is because negative cognitive appraisals and external locus of control are each impacted by their respective baseline status and these baseline constructs likely are correlated. The question is whether there are other meaningful,

unmeasured sources of the correlation between the posttest negative cognitive appraisals and posttest external locus of control that reside in both d_1 and d_3 and whose omission might distort the inferences I care most about (p_1 through p_9). If there is no strong theoretical reason to believe this is the case, then it is not unreasonable to omit the correlation between disturbances d_1 and d_3.

In sum, when modeling an RET you should always carefully consider all pairs of disturbance terms in your model and think about unmeasured confounds that may reside in both terms of the pair to create correlated disturbances. If you are reasonably confident that non-trivial bias will result by ignoring a correlated disturbance, then you should include the correlation in your model rather than ignore it. As noted, it is best to think through these matters *before* conducting your RET and to plan to measure consequential confounders so they can be directly covaried out rather than indirectly dealing with them through correlated disturbances. Introducing correlated disturbances into a model can raise estimation difficulties which I discuss in more depth in the document titled *Dealing with Correlated Disturbances* on the Resources tab of my webpage under Chapter 11. For our RET example, I am going to assume for pedagogical reasons that the measured covariates and baseline variables are sufficient to render the need for correlated disturbances moot.

Working with Latent Variables

Another feature of the RET in Figure 11.2 is its use of a latent variable with three interchangeable indicators of social phobia. As discussed in Chapters 3 and 7, when working with latent variables, we need to assign a metric to it. I use the clinician rating as the reference indicator and pass its metric to the latent variable using the methods discussed in Chapter 7. Re-read that chapter if need-be. The rating ranges from 0 to 6 with clearly demarcated 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.

THE MODEL EQUATIONS

It is helpful for Mplus programming to translate the influence diagram in Figure 11.2 into the implied linear equations but to incorporate the covariates into them. I use *p* notation for the path coefficients and *b* notation for coefficients associated with covariates. I invoke a 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):[1]

$$NCA2 = a_1 + p_1\ T + p_8\ PSS2 + b_1\ BS1 + b_2\ PH1 + b_3\ NCA1 + d_1 \quad [11.1]$$

$$PSS2 = a_2 + p_2\ T + b_4\ BS1 + b_5\ PH1 + b_6\ PSS1 + d_2 \quad [11.2]$$

$$ELC2 = a_3 + p_3\ T + p_9\ PSS2 + b_7\ BS1 + b_8\ PH1 + b_9\ ELC1 + d_3 \quad [11.3]$$

$$LSP3 = a_4 + p_7\ T + p_4\ NCA2 + p_5\ PSS2 + p_6\ ELC2 + b_{10}\ BS1 + b_{11}\ PH1 + b_{12}\ LSP1 + d_4 \quad [11.4]$$

$$CR3 = a_5 + L_1\ LSP3 + e_1 \quad [11.5]$$

$$SPAI3 = a_6 + L_2\ LSP3 + e_2 \quad [11.6]$$

$$SPIN3 = a_7 + L_3\ LSP3 + e_3 \quad [11.7]$$

$$CR1 = a_8 + L_4\ LSP1 + e_4 \quad [11.8]$$

$$SPAI1 = a_9 + L_5\ LSP1 + e_5 \quad [11.9]$$

$$SPIN1 = a_{10} + L_6\ LSP1 + e_6 \quad [11.10]$$

PRELIMINARY ANALYSES

It generally is good practice to check your data relative to assumptions of one's modeling approach. I plan to 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. My analytic strategies also often assume linear relationships between the continuous or many-valued quantitative mediators, covariates, and outcomes. I routinely check the viability of such assumptions by examining scatterplots and smoothers. I also perform checks for outliers and extreme leverages. The *Resources* tab on my webpage provides a document called *Preliminary Analyses for the Social Phobia Example* that describes the form these analyses take and the results of them for the current example. All was in order.

One issue that frequently comes up in regression and SEM analyses is whether one

[1] In causal models where there are causal relationships among two mediators, some methodologists recommend including both baseline mediators as covariates rather than just the target endogenous mediator.

should apply formal statistical tests of model assumptions before embarking on model testing, a common strategy advocated in many statistics books. For example, a preliminary test might take the form of a test of non-normality or of variance heterogeneity. If the preliminary test yields a statistically non-significant result, then one proceeds with the planned analytic strategy that makes the assumption that was tested. If the preliminary test yields a statistically significant result, then one pursues an analytic alternative that either does not make the assumption that was violated or that is robust to violations of it.

Let me state outright that I indeed advocate for exploring your data in depth and that you think long and hard about assumption viability. Having said that, reliance on the above two step approach that uses a preliminary test of assumptions is not as straightforward as many believe (Keselman et al., 2013).

First, many preliminary tests lack statistical power. Without large sample sizes, they can yield non-significant results for testing an assumption violation even when the violation is problematic (see Wilcox, Charlin & Thompson, 1986; Wilcox, 2003). You need to ensure your preliminary test is adequately powered and researchers rarely do so.

Second, in my view, the crucial issue is not whether an assumption is violated (which is what preliminary tests provide perspectives on) but rather the degree to which the assumption is violated. We know that many statistical tests are robust to small violations of their assumptions. What we therefore need to determine is whether the amount of violation present in a given study is consequential. This requires documenting the magnitude of the assumption violation in the sample data and then using margins of error to take sampling error into account when making decisions with respect to that magnitude estimate. It is rare for researchers to do so. Instead, they just rely on the p value from the preliminary test, which tests for the presence of any degree of violation.

Third, many tests of assumptions are based on asymptotic theory and only perform adequately with large sample sizes (Shapiro & Wilk, 1965). However, with large N, preliminary tests tend to detect minor departures from assumption fidelity that may be of little consequence. For tests of non-normality, different tests are sensitive only to certain forms of non-normality, which also can be problematic. For example, some tests are sensitive mostly to skew while others are sensitive mostly to kurtosis.

Fourth, preliminary tests often make assumptions in their own right and may perform poorly when their assumptions are violated. Many tests of variance heterogeneity make normality assumptions and are not robust to violations of normality.

Fifth, using preliminary tests as a screen can change the sampling distribution of key statistics in unpredictable ways. For example, the statistical theory for t tests for comparing two independent means was derived without the idea of first applying a

screening test for normality prior to it. Introducing this step into the process no longer allows all possible random samples of a given size to be part of that sampling distribution. Instead, we are allowing only the sample statistics that have passed the screener to be part of the sampling distribution. The new sampling distribution that results from applying the screener test may no longer be distributed as t, but we still erroneously use the t distribution as the reference distribution for calculating the p value, confidence interval and margin of error.

Although it seems reasonable, for all the reasons I mention above, the practice of conducting preliminary tests is not straightforward. A growing number of statisticians recommend that analysts simply abandon statistical methods that make assumptions of normality and variance homogeneity unless they are confident in assumption viability based on theory or prior research. This is why, for example, a robust version of maximum likelihood estimation often is preferred to the more traditional maximum likelihood version of SEM when evaluating a model; one does not have to worry so much about the assumptions of non-normality and variance homogeneity in the first place. Instead of conducting flawed and underpowered preliminary tests and altering analytic strategies based on the results of those preliminary tests, the preferred approach is to use methods that do not make those assumptions in the first place and do not lead to sacrifices in statistical power (Keselman et al., 2008; Wilcox, 2017). Cases can occur where defaulting to robust analytic strategies may result in some loss of statistical power and/or somewhat larger margins of error. However, in the long run, the argument goes, the use of robust methods often will result in better Type I error protection, increased power to detect effects, and confidence intervals that more accurately reflect the desired probability coverage as compared to the flawed two step strategy (Wilcox, 1998).

In sum, I personally view preliminary tests with some skepticism. If I apply one and obtain a statistically significant result that suggests assumption violation, I am left wondering (a) whether I can trust the preliminary test given the assumptions it makes, and (b) whether the degree of violation that is operating reaches a level that I have to worry about. If I obtain non-significant results that are consistent with the absence of assumption violations, I wonder about (a) whether there was sufficient power in the preliminary test to detect meaningful levels of assumption violation, and (b) how well the preliminary test performs in scenarios that map onto my sample size. For example, some normality tests perform badly for sample sizes less than 400. Independent of the above, I also worry about the impact on the sampling distributions of my statistical tests by making their application contingent on the results of an imperfect “screening” test.

I make it a point to explore data preliminarily to understand analytic complications that I need to be wary of. But this typically goes beyond reliance on p values associated

with preliminary tests. When I conduct preliminary analyses and see patterns that are consistent with assumptions of the test I intend to apply, I gain a sense of reassurance. When I see patterns that suggest analytic complications, I try to deal with them. My own bias is to use robust methods of analysis that do not require the assumptions in the first place. For example, the MLR estimation method in Mplus is robust to many forms of non-normality and to variance heterogeneity. Traditional maximum likelihood analysis, by contrast, assumes multivariate normality among the indicators, so I only use it in special circumstances.

TRADITIONAL FULL INFORMATION SEM ANALYSIS

In this section, I conduct traditional FISEM analyses with robust maximum likelihood using Mplus. Mplus relies on syntax that is written and executed using the Mplus interface. My website provides links to assorted programming resources and a general tutorial on Mplus programming. I highlight here the Mplus syntax for the social phobia RET. The syntax is in Table 11.1. I numbered each line for referencing but the line numbers are not part of Mplus syntax. The numbers should be excluded when you write your Mplus programs. I provide a video describing how to program the model that you can watch if you prefer that form of learning. Here is the video link: Mplus Syntax. [*If you are not reading this pdf in a browser, then just left click the link. If you are reading it from within Chrome, right click the link and choose to open the link in a new window; if reading it from within Safari, hold down the command key while clicking the link. If reading a printed copy, see the link on the Resources tab of my webpage for Chapter 11*.]

Table 11.1: Mplus Syntax for Social Phobia Example

```
1. TITLE: EXAMPLE CHAPTER 11 ;
2. DATA: FILE IS c:\mplus\ret\chap11M.txt ;
3. VARIABLE:
4. NAMES ARE ID CR1 SPAI1 SPIN1 CR3 SPAI3 SPIN3
5. NEGAPP2 PSKILLS2 EXTERN2 NEGAPP1 PSKILLS1 EXTERN1
6. HYPER SEX TREAT ;
7. USEVARIABLES ARE CR1 SPAI1 SPIN1 CR3 SPAI3 SPIN3
8. NEGAPP2 PSKILLS2 EXTERN2 NEGAPP1 PSKILLS1 EXTERN1
9. HYPER SEX TREAT ;
10. MISSING ARE ALL (-9999) ;
11. ANALYSIS:
12. ESTIMATOR = MLR ;  !Robust maximum likelihood
13. MODEL:
14. !Specify latent variables
15.    LSP1 BY CR1 SPAI1 SPIN1 ;
16.    LSP3 BY CR3 SPAI3 SPIN3 ;
```

```
17. !Specify equations
18. LSP3 ON LSP1 NEGAPP2 PSKILLS2 EXTERN2 TREAT SEX (b10 p4-p7 b11) ;
19. LSP3 ON HYPER (b12) ;
20. NEGAPP2 ON TREAT HYPER SEX NEGAPP1 PSKILLS2 (p1 b1-b3 p8) ;
21. PSKILLS2 ON TREAT HYPER SEX PSKILLS1 (p2 b4-b6) ;
22. EXTERN2 ON TREAT HYPER SEX EXTERN1 PSKILLS2 (p3 b7-b9 p9) ;
23. !Specify correlations of latent variable with exogenous variables
24. LSP1 WITH NEGAPP1 PSKILLS1 EXTERN1 TREAT SEX HYPER ;
25. MODEL INDIRECT:
26. LSP3 IND TREAT ;
27. LSP3 IND PSKILLS2 ;
28. NEGAPP2 IND TREAT ;
29. EXTERN2 IND TREAT ;
30. OUTPUT:
31. SAMP STANDARDIZED(STDYX) MOD(ALL 4) RESIDUAL CINTERVAL TECH4 ;
```

Mplus uses keyword commands that are followed by a colon. Within the commands are subcommands that are terminated by a semi-colon. A given line cannot be longer than 90 characters, including spaces. If your subcommand has more than 90 characters, enter a carriage return before you reach 90 characters and continue typing on the following line. Mplus is **not** case sensitive. Also, you can have as many spaces as you like between words. Line 1 is a title line. Line 2 specifies the `DATA` command and tells Mplus where to find the data file. I use .txt for the file tag, but you can use any tag (e.g., .dat). The file typically is an ASCII file in free format, where the values for each variable are separated by a delimiter, usually a blank, a tab, or a comma. Other formats are available but I do not consider them here; see the Mplus users guide. The numbers in the data file are arranged so that the first person's scores on each variable appear first, followed by the second person's scores, and so on until the last person's scores are listed. Exported ASCII files from most software is usually compatible with Mplus. Do not write variable names on the first line of the file. Only numbers are valid in the data file.

Line 3 is the `VARIABLE` command and tells Mplus information about the variables will come next. Line 4 contains the subcommand `NAMES ARE`, followed by the names of the variables in the order they appear in the data file. The names are separated by spaces and must be 8 characters or less. I used three lines in this case to avoid the 90 character line limit. You can have as many spaces as you want between variable names. Lines 7 to 9 specify the subset of variables from those on the `NAMES ARE` command that are to be used in the analysis. `USEVARIABLES` is the subcommand, followed by the variable list. In this case, I use every variable except the `ID` variable. If you are going to use all of the variables in the `NAMES ARE` list, then the `USEVARIABLES` command can be omitted. Line 10 tells Mplus that all variables have missing values of -9999. When Mplus encounters a -9999 in the data, it treats it as missing. You can use any value to signify missing data.

Line 11 is the ANALYSIS command and indicates details of the desired analysis are to follow. The subcommand ESTIMATOR = MLR invokes the Huber-White robust estimator. I include an exclamation point on this line, which signifies a comment; all text on a line after a ! is ignored by Mplus, but the 90 character limit still applies. Line 13 is the MODEL command and tells Mplus I will now specify the model. Lines 15 and 16 specify the latent variables using the keyword BY. The latent variable name is listed on the left of BY and the indicators are listed to the right of BY. The latent variable name can be of your choosing but it cannot exceed 8 characters. I use LSP to reflect "latent social phobia," followed by a number to indicate the time of assessment. Mplus assumes the first indicator listed after BY is the reference indicator whose loading is fixed to equal 1.0.

By default, Mplus estimates the measurement intercepts for each indicator of a latent variable and it fixes the underlying latent variable mean and latent intercept to zero. It turns out that this default method of handling factor means and intercepts does not affect the results we care about, at least for the current example. In later chapters, I will override the defaults and use the factor means and intercepts.

Lines 18 to 22 specify the model equations using the ON keyword. The endogenous variable is listed to the left of the ON and all the predictors of the endogenous variable are listed to the right. The ON keyword tells Mplus to estimate a path coefficient for each of the predictors. The four equations I program are described in the Model Equations section presented above. After the variables listed using the ON keyword, I added notation to assign labels to the path coefficients. You add labels for a parameter by adding the label in parentheses right after the parameter and before the semi-colon terminator. You can use any labels you want up to 8 characters. I do not make use of the labels in this chapter but I do use them in some of the supplementary materials for this chapter and in future chapters, so I introduce the practice here. Consider the ON statement for PSKILLS2 on line 21. This ON statement has four predictors, TREAT HYPER SEX and PSKILLS1. all listed on the same line. Mplus will look for four labels within the parentheses given there are four predictors. To match the path labels I used in Figure 11.2 and the model equations, I use the labels (p2 b4 b5 b6). Each coefficient label is separated by at least one space. Mplus offers a shorthand if the labels are the same except for a trailing number that increments by one. Instead of writing out b4 b5 b6, I can use b4-b6. This assigns sequential labels beginning with 4 and extending through 6, in this case, b4 b5 b6. I use this shorthand in lines 18 to 22. The assignment of labels is optional.

Another feature of Mplus is shown in Lines 18 and 19. The two lines represent a single equation. Line 19 was close to the per line character limit of Mplus. Normally, I would enter a carriage return and continue the entry. However, because I am using labels, I can't do this. I specify part of the equation with labels on line 18. Then I specify an

equation on line 19 with the same outcome and the predictors I would like to add to the first equation with their labels. During analysis, Mplus will merge the two equations.

By default, Mplus treats exogenous variables as fixed predictors and assumes they are correlated with one another, although the correlations are not formal parameters in the model. An exception is when there are exogenous latent variables, in which case, you need to explicitly tell Mplus to estimate the correlations between the exogenous latent variables and the other exogenous variables; otherwise Mplus assumes the correlations are zero. This is done using the `WITH` command in which the variable named to the left of `WITH` is correlated with all variables to the right of `WITH`, per line 24. When I introduce this line, the correlations between all exogenous to the right of `WITH` are formally parameterized in the model as well, bringing with them underlying distributional assumptions. In general, if one exogenous predictor is treated as random rather than fixed, statistical theory dictates that all of the predictors be treated as random. This usually has little consequence for results, at least for continuous outcomes.

Line 25 tells Mplus I want a detailed analysis of mediated (indirect) effects. Line 26 asks for such an analysis for the outcome variable listed to the left of `IND`, in this case LSP3, and the distal determinant to the right of `IND`, in this case `TREAT`. I also include an indirect effect analysis from `PSKILLS2` to `LSP3`, `TREAT` to `NEGAPP2`, and `TREAT` to `EXTERN2` on lines 27 to 29, for reasons I explain later. Lines 30 and 31 tell Mplus what output information I want beyond the defaults. `SAMP` requests sample descriptive statistics; `STANDARDIZED(STDYX)` requests standardized coefficients in addition to unstandardized coefficients; `MOD(ALL 4)` asks Mplus to show all modification indices whose value is equal to or greater than 4; `RESIDUAL` asks Mplus for residual analyses of the difference between predicted and observed covariances; `CINTERVAL` asks Mplus for traditional confidence intervals; and `TECH4` asks for technical output I explain later.

I provide an annotated copy of the full Mplus output on the Resources tab on my webpage (see Chapter 11). Here, I consider first the output for model fit. I then describe results for (a) the total effect of the treatment on social phobia, (b) the relationship between the targeted mediators and the outcome, and (c) the effect of the program on the targeted mediators.

Results of the Analysis: Model Fit

The RET measured 15 variables that yielded a 15X15 covariance matrix. I hypothesized that the covariance patterns in this matrix are due to the causal dynamics in Figure 11.2 plus the covariates. The model makes predictions about how the observed covariances should pattern themselves. The question is whether the covariances pattern themselves in a way that is consistent with model predictions. If not, I reject the model as viable. If the

observed covariances pattern themselves as predicted, I have increased confidence in the model. Note that if the patterning of data is consistent with model predictions, this does not prove the model is correct. Rather, it typically increases confidence in the model.

Examination of the global fit indices is a first step in evaluation of model-data correspondence. The fit statistics I rely on appear on the output as follows:[2]

```
MODEL FIT INFORMATION

Chi-Square Test of Model Fit

          Value                             50.341*
          Degrees of Freedom                    57
          P-Value                           0.7213
          Scaling Correction Factor         1.0094
            for MLR

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.000
          90 Percent C.I.     0.000  0.026
          Probability RMSEA <= .05           1.000
CFI/TLI

          CFI                                1.000
          TLI                                1.000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.017
```

The chi square test of perfect model fit in the population was statistically non-significant (chi square = 50.34 with 57 degrees of freedom, $p < 0.73$), which is consistent with the proposition that the data are reasonably in accord with the model. The RMSEA was less than 0.001, which is consistent with good model fit. Although the output reports the value as being 0.000, this sometimes occurs because Mplus only shows results to three decimals. The upper limit of the 90% confidence interval is 0.026, which also is consistent with a reasonable model fit. The CFI is 1.00 and the standardized RMR was 0.017, which both suggest reasonable model fit.

Despite the above, I am not prepared to declare that the data are consistent with the model until I also examine localized fit indices. Mplus reports for each cell of the correlation matrix the values of the observed correlations minus the predicted

[2] I edit the output to save space. A warning message appears on the output about a non-positive definite first order product matrix. This warning can be ignored in this case; see the document in Chapter 11 on the Resources tab of my webpage.

correlations. All of these should be near zero. I routinely examine this "residual matrix" to ensure the entries are near zero. The matrix appears in the output section called `RESIDUAL OUTPUT` in the subsection called `Residuals for Correlations`:

```
Residuals for Correlations
             CR1           SPAI1         SPIN1         CR3           SPAI3
             ________      ________      ________      ________      ________
 CR1           0.000
 SPAI1         0.000         0.000
 SPIN1         0.002        -0.002         0.000
 CR3           0.020         0.011        -0.020         0.000
 SPAI3         0.003         0.002        -0.027         0.000         0.000
 SPIN3         0.030         0.011         0.008        -0.002         0.002
 NEGAPP2       0.056        -0.008        -0.015         0.000        -0.012
 PSKILLS2     -0.049         0.000         0.025        -0.009        -0.001
 EXTERN2       0.053         0.024         0.023        -0.004        -0.017
 NEGAPP1      -0.001         0.005        -0.004        -0.012        -0.013
 PSKILLS1     -0.027        -0.007         0.031        -0.021         0.013
 EXTERN1      -0.021         0.021         0.000         0.081         0.024
 HYPER        -0.003        -0.008         0.010        -0.002        -0.008
 SEX           0.014         0.009        -0.021         0.013        -0.022
 TREAT        -0.020         0.008         0.013         0.002         0.007


Residuals for Correlations
             SPIN3         NEGAPP2       PSKILLS2      EXTERN2       NEGAPP1
             ________      ________      ________      ________      ________
 SPIN3         0.000
 NEGAPP2       0.005         0.000
 PSKILLS2      0.011         0.008         0.000
 EXTERN2       0.019        -0.016         0.000         0.000
 NEGAPP1      -0.015        -0.019         0.046        -0.058         0.000
 PSKILLS1      0.014        -0.019         0.000         0.010         0.000
 EXTERN1       0.042         0.066         0.001         0.000         0.000
 HYPER         0.007         0.001         0.000         0.000         0.000
 SEX           0.004         0.000         0.000         0.000         0.000
 TREAT        -0.007        -0.004         0.000         0.000         0.000


             PSKILLS1       EXTERN1         HYPER         SEX           TREAT
             ________      ________      ________      ________      ________
 PSKILLS1      0.000
 EXTERN1       0.000         0.000
 HYPER         0.000         0.000         0.000
 SEX           0.000         0.000         0.000         0.000
 TREAT         0.000         0.000         0.000         0.000         0.000
```

Values that are exactly zero likely are from tautological predictions per my discussion in Chapter 7. There are no large disparities in the present case. The table is complemented by examining per cell significance tests of the difference between each predicted and observed variance/covariance. These tests take the form of z tests and are in the section called `Standardized Residuals (z-scores) for Covariances`. If the

absolute value of a cell entry is larger than 1.96, the null hypothesis of no difference between the predicted and observed covariances should be rejected. Here is the output:

```
Standardized Residuals (z-scores) for Covariances

              CR1           SPAI1         SPIN1         CR3           SPAI3
              _____         _____         _____         _____         _____
 CR1           0.000
 SPAI1        -0.108         0.000
 SPIN1         0.714        -1.445         0.000
 CR3           0.704         0.357        -0.667        -0.099
 SPAI3         0.102         0.069        -0.832        -0.090        -0.079
 SPIN3         1.084         0.365         0.364        -0.299         0.256
 NEGAPP2       1.348        -0.187        -0.457        -0.300        -1.246
 PSKILLS2     -1.365         0.001         0.710        -0.783        -0.093
 EXTERN2       1.059         0.449         0.463        -0.238        -0.893
 NEGAPP1      -0.028         0.216        -0.209        -0.340        -0.407
 PSKILLS1     -1.812        -0.319         1.595        -0.635         0.382
 EXTERN1      -1.130         0.998         0.028         2.266         0.654
 HYPER        -0.202        -0.317         0.458        -0.118        -0.462
 SEX           0.689         0.430        -1.171         0.770        -1.183
 TREAT        -1.069         0.366         0.728         0.208         0.505


Standardized Residuals (z-scores) for Covariances

              SPIN3         NEGAPP2       PSKILLS2      EXTERN2       NEGAPP1
              _____         _____         _____         _____         _____
 SPIN3        -0.114
 NEGAPP2       0.126       999.000
 PSKILLS2      0.683         1.327         0.000
 EXTERN2       0.969        -0.622         0.031        -0.031
 NEGAPP1      -0.473        -1.671         1.377        -1.158         0.000
 PSKILLS1      0.428        -0.560         0.000         0.227         0.000
 EXTERN1       1.075         2.134         0.032        -0.026         0.000
 HYPER         0.339         0.000         0.000         0.000         0.000
 SEX           0.310         0.000         0.000         0.000         0.000
 TREAT        -0.565         0.000         0.000         0.000         0.000


              PSKILLS1      EXTERN1       HYPER         SEX           TREAT
              _____         _____         _____         _____         _____
 PSKILLS1      0.000
 EXTERN1       0.000         0.000
 HYPER         0.000         0.000         0.000
 SEX           0.000         0.000         0.000         0.000
 TREAT         0.000         0.000         0.000         0.000         0.000
```

Each entry is the variance/covariance difference divided by the estimated standard error of the difference. The diagonal elements test the differences between the predicted and observed variances and the off diagonals test the differences between the predicted and observed covariances. On occasion, values of 999 occur. This happens when Mplus is not

able to calculate the z test, usually because of a negative standard error. In these cases, the statistic is ignored. When it does occur, some researchers examine the corresponding entries in a matrix called `Normalized Residuals for Covariances` for the cell where the 999 occurred. This latter matrix takes the same form as the standardized residuals but the variance/covariance differences are divided by the estimated standard errors of the observed variances/covariances rather than the standard error of the difference between the observed and predicted variances/covariances. Absolute values larger than 1.96 suggest statistically significant disparities, but these tests are on weaker statistical grounds than the standardized residuals that use the more appropriate standard error.

When evaluating these significance tests, one must take into account the large number of tests performed. In the present model there were 120 such significance tests, so 5 or 6 could be statistically significant by chance alone. Because I generated the data from a population model that perfectly mapped onto the tested model, I know for a fact that any significant results are chance. In the present data, only two absolute standardized z values greater than 1.96 occurred. In practice, I could apply the False Discovery Rate (FDR) method to control for multiplicity across the 120 tests (see Chapter 6), but some methodologists argue that doing so is too conservative. Another possibility is to use the FDR method but only for contrasts that are not mathematical tautologies. There were 88 such contrasts. After considering possible chance results using the program for FDR controls on my website, there did not appear to be significant disparities. Some methodologists suggest using forms of control for multiplicity other than the FDR method (e.g., a Holm modified Bonferroni method), but I often find they reduce statistical power to detect misspecification too much to be satisfactory.

Another localized test of model fit is to examine the modification indices to determine if there are points of stress in the model relative to omitted parameters. Issues of multiplicity also must be taken into account for modification indices. Indices larger than 4.0 are of interest because adding the parameter would likely yield a statistically significant result for that parameter. Here is the relevant output (edited to save space) :

```
MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index     4.000

                                  M.I. E.P.C. Std E.P.C. StdYX E.P.C.
ON Statements

CR1      ON NEGAPP2               5.076     0.062      0.062        0.067
CR1      ON PSKILLS2              4.684    -0.061     -0.061       -0.064
CR3      ON EXTERN1               4.281     0.160      0.160        0.051
NEGAPP2  ON EXTERN1               5.046     0.137      0.137        0.075
```

```
WITH Statements

EXTERN1  WITH CR3                  4.723     0.029      0.029        0.131
EXTERN1  WITH NEGAPP2              4.811     0.022      0.022        0.111
HYPER    WITH NEGAPP2              4.666    -0.066     -0.066       -0.331
```

Seven modification indices (under the heading M.I.) were larger than the critical value of 4.0. The statistically significant modifications make little conceptual sense. For example, the first index listed suggests I regress the clinician report of social phobia at baseline onto the patient's negative appraisals at posttest, which violates the time ordering of causal relationships. The first index in the `WITH` category suggests I correlate baseline external locus of control with the clinician ratings at the three-month follow-up, which makes no conceptual sense. Given this and the large number of modification indices evaluated, it is likely these elevated indices reflect chance (in fact, I know this is the case because I generated the sample data from a population that maps onto the tested model). To determine how many modification indices that Mplus computes, change the command on the output line (line 31) to `MOD(ALL 0)` and re-run the program. The new output will show all of the modification indices, not just those greater than 4.0. When I did so, there were 272 modification indices. Again, I could apply the FDR method to control for multiplicity across these contrasts. I conclude there are no meaningful points of stress in model fit associated with omitted parameters.

On the output for modification indices, `E.P.C.` stands for **expected parameter change** and is the estimated value that the parameter would take on if it were to be added to the model given that it is zero to start with. `StdYX E.P.C.`is the same concept but expressed in a fully standardized metric. For a `WITH` statement, `StdYX E.P.C` is a correlation coefficient. For an `ON` statement, `StdYX E.P.C` is a standardized path coefficient and `E.P.C.` is an unstandardized path coefficient. For a `BY` statement, `StdYX E.P.C` is a standardized factor loading. As discussed in Chapter 7, all of these statistics can be used to judge the likely effect size of the parameter if it were to be added to the model. Even if a modification index is statistically significant, the effect size might be sufficiently trivial that one decides to leave the parameter out anyway.

For further discussion of the localized fit indices, see Chapter 7 and Kline (2024). In the current case, both global and localized fit indices suggest the data are reasonably consistent with model predictions. It makes sense to interpret the parameter estimates.

Results of the Analysis: Evaluation of Measurement Model

Prior to formal substantive interpretations, I examine the measurement model for the latent variables to gain a sense of the psychometric properties of the measures used to

assess them. The parameters of interest typically are the standardized factor loadings and the standardized measurement error variances. Here is the output for the standardized error variances taken from the section labeled `STDYX Standardization` and the subsection called `Residual Variances`:

```
                                                    Two-Tailed
                    Estimate       S.E.     Est./S.E.    P-Value
Residual Variances
    CR1                0.188      0.024      7.896      0.000
    SPAI1              0.232      0.030      7.788      0.000
    SPIN1              0.183      0.028      6.449      0.000
    CR3                0.141      0.018      8.019      0.000
    SPAI3              0.168      0.020      8.379      0.000
    SPIN3              0.116      0.017      6.703      0.000
```

The column labeled `Estimate` is the standardized parameter estimate. Given the use of interchangeable indicators, the entries can be thought of as representing the proportion of variation in each measure that is due to random noise (unreliability). One minus these values is the estimated reliability of the measure. For example, the reliability of the clinician ratings at follow-up is 1-0.14 = 0.86. The column labeled S.E. is the estimated standard error of the unreliability estimate. A back-of-the-envelope estimate of the 95% margin of error (MOE) for the estimate is to double the value of the estimated standard error. The rationale is that the estimated standard error is an index of the standard deviation of the parameter from one random sample to the next in a sampling distribution, so it directly reflects sampling error. The estimated standard error is doubled to mimic a 95% confidence interval that multiplies the estimated standard error by a critical value of 1.96, assuming the N is not too small. Using this heuristic, the estimated unreliability of the clinician ratings at the follow-up is 0.14 ±0.04. For SPAI it is 0.17 ±0.04. For the SPIN, it is 0.12 ±0.03. I often use this approach in this book.

The column labeled `Est./S.E.` is the parameter estimate divided by its estimated standard error. It is called a z value or a **critical ratio**. It is a significance test of the null hypothesis that the standardized error variance of measure is zero, i.e., that there is perfect reliability. A critical ratio greater than 1.96 is statistically significant. The last column is the two tailed p value for the critical ratio. These tests usually are not of interest because they almost always yield $p < 0.05$; it is rare for indicators of a latent variable to have perfect reliability. The tests for standardized coefficients also are only approximate. More accurate tests are available from bootstrapped solutions.

A more precise MOE for the standardized coefficients uses the applicable 95% confidence interval. This is in the output section called `CONFIDENCE INTERVALS OF STANDARDIZED MODEL RESULTS` in the subsection called `Residual Variances`. Here is

the output for the three follow-up social phobia indicators:

	Lower .5%	**Lower 2.5%**	Lower 5%	**Estimate**	Upper 5%	**Upper 2.5%**	Upper .5%
CR3	0.095	**0.106**	0.112	**0.141**	0.170	**0.175**	0.186
SPAI3	0.117	**0.129**	0.135	**0.168**	0.202	**0.208**	0.220
SPIN3	0.072	**0.082**	0.088	**0.116**	0.145	**0.150**	0.161

I bolded the entries of interest, but the bolding does not occur on Mplus output. The column labeled `Estimate` is the parameter estimate. The columns to the right of it are the upper limits of confidence intervals and the columns to the left are the lower limits. The lower and upper 5% entries correspond to a 90% confidence interval, the lower and upper 2.5% entries correspond to a 95% confidence interval, and the lower and upper .5% entries correspond to a 99% confidence interval. The 95% confidence interval for `CR3` is 0.106 to 0.175. The 95% upper MOE for `CR3` is 0.175 - .141 = 0.03 and the 95% lower MOE is 0.106 - .141 = -.04. If I use the larger of the absolute value of the two limit differences to summarize the result, the MOE is ±0.04. Note this is the same as the estimate based on the heuristic "double the standard error." Some methodologists prefer to report both the lower and upper MOE separately to be more precise.

The above MOEs were based on robust maximum likelihood estimation coupled with a method that assumes symmetric confidence intervals about the parameter estimate. Although this often is reasonable for unstandardized parameters, standardized parameters sometimes have asymmetric confidence intervals. The asymmetric intervals can be obtained by re-running the program with bootstrapping; change the estimator from MLR to ML on line 12 of Table 11.1, and add the text `BOOTSTRAP=5000` to it, as follows:

```
ESTIMATOR = ML ; BOOTSRAP = 5000 ;
```

The number 5000 specifies the number of bootstrap replicates to use. Then, on the output line (line 32), change the confidence interval statement to read

```
CINTERVAL(BOOTSTRAP)
```

As well, I remove the `MOD(ALL 4)` term on the output line (line 32) because Mplus does not permit modification indices with bootstrapping. The output is identical in format to when I used MLR, but all standard errors, p values, and confidence intervals are bootstrapped. Here are the confidence intervals for the standardized error variances:

	Lower .5%	**Lower 2.5%**	Lower 5%	**Estimate**	Upper 5%	**Upper 2.5%**	Upper .5%
CR3	0.098	**0.109**	0.113	**0.141**	0.171	**0.177**	0.192
SPAI3	0.122	**0.132**	0.137	**0.168**	0.203	**0.212**	0.227
SPIN3	0.075	**0.083**	0.088	**0.116**	0.146	**0.152**	0.165

The values are close to those from the MLR analysis and the amount of asymmetry is trivial. Reporting the MLR estimates or even the "double the standard error" estimates is not unreasonable in this case.

Here are the results for the standardized factor loadings from the MLR analysis from the output section labeled `STDYX Standardization`:

	Estimate	S.E.	Est./S.E.	P-Value
LSP1 BY				
CR1	0.901	0.013	68.030	0.000
SPAI1	0.876	0.017	51.514	0.000
SPIN1	0.904	0.016	57.622	0.000
LSP3 BY				
CR3	0.927	0.009	97.966	0.000
SPAI3	0.912	0.011	82.725	0.000
SPIN3	0.940	0.009	102.024	0.000

For `CR3` and using the doubled standard error MOE heuristic, the standardized loading for `CR3` is 0.93 ±0.02, for `SPAI3` it is 0.91 ±0.02 and for `SPIN3` it is 0.94 ±0.02. These estimates are standardized path coefficients; for example, for every one standard deviation that `LSP3` increases, `CR3` is predicted to increase by 0.93 standard deviations.

An interesting property of the standardized factor loadings is that if you multiply the loading for one variable by the loading for another variable, the result is the predicted correlation between the two variables. For example, the (predicted) correlation between `CR3` and `SPAI3` is (0.927)(0.912) = 0.84. This property only holds for variables that are indicators of the same latent variable and when there is no correlated error or cross-loadings. Another interesting property is that if you square the factor loading and subtract it from 1.0, you will obtain the standardized error variance. For `CR3`, $1 - 0.927^2$ is .14, which is the standardized error variance for `CR3`. Knowledge of this property is useful if investigators fail to report measure unreliability in their research reports. It follows from this property that the square of the factor loading is the reliability estimate; for `CR3`, for example, the reliability is estimated as (0.927)(0.927) =0.86.

In sum, the measures used to assess social phobia at baseline and posttreatment seem to have reasonable reliability. Note that this does not mean they are valid. Rather, it means they are relatively free of random error. However, the fact that they are highly correlated with one another as well is evidence for their convergent validity.

Another psychometric issue that I can address in the data is whether the social phobia indicators have properties of measurement invariance, as discussed in Chapter 3. For latent variables, SEM has the capability of evaluating non-invariance of measurement

intercepts and factor loadings across groups and time. The presence of measurement non-invariance can undermine the interpretation of group differences in means and path coefficients, as I discussed in Chapter 3. I used the methods described in the primer for Chapter 3 to test for measurement non-invariance in the current RET. I found no evidence for consequential measurement non-invariance.[3]

Results of the Analysis: Total Effect of the Program on the Outcome

The first substantive question is whether the program affected the outcome and by how much. I first need to set my meaningfulness standards in order to evaluate the total effect. Using the latitude framework from Chapters 2 and 10, suppose that after consultation with clinicians and relevant staff, a consensus was reached that an average change of -1.0 units on the latent variable metric for the clinician rating (CR3) represents meaningful change, i.e., -1.0 is the lower bound of the latitude of meaningfulness. Suppose it also was agreed that changes between –0.50 and 0.50 are deemed to be inconsequential, i.e., the absolute value of 0.50 defines the boundaries for concluding no effect. With these standards in mind, I examine the relevant statistics for the total effect,

In FISEM, the total effect is *not* determined by directly comparing mean outcome values for the treatment and control groups. Rather, the total effect is parameterized (a) by assuming the tested causal model is correct, and then (b) combining all of the relevant estimated path coefficients that lead, directly or indirectly, from the treatment condition to the outcome, per my discussion of multiplicative rules for combining coefficients in Chapters 5 and 7. As such, the total effect of the treatment is model-defined and evaluated as such. Fortunately, Mplus does all the tedious calculations for you.

The estimated value of the total effect of the treatment on the outcome is in the output section labeled `TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS`. Here is the relevant output:

```
                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.     P-Value

Effects from TREAT to LSP3

  Total             -1.758      0.104    -16.855      0.000
  Total indirect    -1.270      0.121    -10.463      0.000
```

[3] I, of course, also routinely evaluate the psychometrics of the multi-item single indicator measures per the methods I presented in Chapter 3 but do not do so here in the interest of space.

The outcome is the post treatment latent social phobia variable that is scaled continuously on the 0 to 5 metric of the clinician rating and that uses scale adverb modifiers 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 estimated mean difference between the treatment and control groups is -1.76 ±0.21 ($z = 16.86$, $p < 0.05$). The negative coefficient means a larger value (the control group mean) was subtracted from a smaller value (the treatment group mean), assuming dummy coding. On average, individuals in the treatment condition improved by -1.76 units relative to the control group after adjusting for baseline social phobia and the covariates. If individuals had scores near 4 on the metric prior to treatment ("quite social phobic, quite disabling"), they were likely to have scores near 4 - 1.76 = 2.24 (just above "moderate social phobia, somewhat disabling") at follow-up. If individuals had scores near 3 on the metric prior to treatment ("social phobic, moderately disabling"), they were likely to have scores near 3 - 1.76 = 1.24 (just above "mild social phobia, not disabling") at follow-up. And so on.

Because the total effect is a non-linear combination of multiple path coefficients in the model, it usually is best to use percentile bootstrapping to estimate its statistical significance and margins of error. I implemented bootstrapping using the altered syntax described earlier and the results were quite close to the MLR method. The total effect was -1.76 ($z = 16.51$, $p < 0.05$, lower MOE = -0.21, upper MOE = 0.21).

Meaningfulness Standard for the Program Total Effect

I next compare the observed total effect and its confidence limits to the meaningfulness standards using the latitude framework from Chapters 2 and 10. Recall that the agreed upon standard for the latent CR3 metric was a change of -1.0 units and the standard for a trivial effect was any absolute change of 0.50 or less. The mean difference of -1.76 exceeded the meaningfulness standard of -1. The 95% confidence interval for the total effect was -1.97 to -1.55. Note that even the upper limit of this interval (-1.55) is less than the meaningfulness standard of -1.0. This result leads me to confidently conclude that the program did indeed have a meaningful effect on the outcome even after taking into account sampling error. It turns out the mean clinician rating for the control group at posttest was 3.12, which is near the anchor of 3 = social phobic, moderately disabling. Subtracting 1.76 from this reference value yields a posttest mean of 1.36, a rating that is between the metric points of 1 = mild social phobia, not disabling and 2 = moderate social phobia somewhat disabling, with the mean closer to the former.

Standardized Effect Size Indices for the Total Effect

The conversion of the total effect to a probability of exception to the rule (P_E), a Cohen's *d*, or a percent of explained variance using FISEM is not straightforward for a total effect. This is because the total effect is a complex function of the many paths that connect the treatment condition dummy variable, `TREAT`, to the outcome, `LSP3`. The critical ratio for the total effect does not follow simple OLS regression rules and the calculation of its standard error is complex. The regularities for calculating standardized effect sizes that I described in Chapter 10 do not apply in this case. One solution is to work outside the FISEM framework using a simplified estimation method. For example, an ANCOVA-like model would estimate the total effect by regressing `LSP3` onto the treatment condition and the relevant covariates of biological sex, hypercriticism by parents, and the baseline latent social phobia variable. I can then use the methods described in Chapter 10 to calculate effect size analogs. I describe the process for the social phobia example in the document *Effect Size for Social Phobia Example* on the Resources tab for Chapter 11 on my website. When I applied the simplified analysis and calculated a probability of exceptions to the rule, I found the approximate value for P_E wase 0.08. The general rule formed based on the data was that people who participate in the program tend to have lower social phobia than people who do not participate in the program. The level of exceptions to this rule, expressed as the P_E converted to a percent, is 8%. For random draws of pairs of people, one from each group, about 8% of the time there will be exceptions to the rule, holding constant baseline social phobia, biological sex, and the person's history of parental hypercriticism. Cohen's d was -2.02 and the unique explained variable accounted for by the treatment condition was 0.39.

In sum, does the program bring about meaningful change in social phobia? Yes it does. Can the program be improved? Yes, it can as elaborated below.

Results of the Analysis: Effects of the Program on the Mediators

The next question is how successful the program is in changing the targeted mediators. I first derive meaningfulness standards for each mediator and then consider the results.

Meaningfulness Standards for Program Effect on Mediators

Each of the three mediators is measured on the same -3 to +3 metric. To define a meaningfulness standard for program effects on the mediators, I can use the strategy discussed in Chapter 10 in conjunction with the program on my website called *Effect size standards*. Suppose instead I engaged in extensive discussions with program staff and clients and conclude that a reasonably meaningful effect for a mediator is ¾ of a scale

unit or 0.75. For perceived social skills this effect should be in a positive direction; for negative cognitive appraisals and external locus of control, this should be in a negative direction.

Effect of Program on Perceived Social Skills

For perceived social skills, the information for the program effect on PSKILLS2 is in the section on MODEL RESULTS. Here is the relevant output:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PSKILLS2 ON				
TREAT	1.173	0.050	23.638	0.000
HYPER	0.022	0.061	0.363	0.716
SEX	0.045	0.049	0.912	0.362
PSKILLS1	0.508	0.059	8.618	0.000

Of interest is the coefficient from TREAT to PSKILLS2. The program increased mean perceptions of social skills by 1.17 ±0.10 units on the -3 to +3 disagree-agree metric of PSKILLS2 relative to the control group ($z = 23.64$, $p < 0.05$), holding constant the other predictors in the equation. The meaningfulness standard for T→PSKILLS2 is that the coefficient should be ≥0.75. This was the case, but I also need to take into account sampling error when making a conclusion. The 95% confidence interval for the T→PSKILLS2 link was 1.07 to 1.27. Because the lower limit of the 95% confidence interval exceeds the meaningfulness standard, I can conclude the program had a meaningful effect on perceived social skills after taking into account sampling error.

For standardized effect size indices of the effect of the program on perceived social skills, I can calculate the probability of exceptions to the rule using the program for binary predictors on my website. The general rule is that people who participate in the intervention tend to have higher perceived social skills than people who do not participate in the intervention. Using the program on my webpage, I find that the level of exceptions to this rule is about 0.03 or 3%: For random draws of pairs of people, one from the treatment group and one from the control group, about 3% of the time, the person from the control group will have *higher* perceptions of his or her social skills than the person from the treatment group, holding constant baseline perceptions of social skills, biological sex, and the person's history of parental hypercriticism. If you want to calculate Cohen's *d* or a percent of explained variance for T→PSKILLS2, see the document for effect sizes on my webpage in the Resources tab for Chapter 11.

Effect of Program on Negative Cognitive Appraisals

For the negative cognitive appraisals mediator, Figure 11.2 indicates there are two ways by which the program affects it. First, there is a direct effect of the treatment on negative cognitive appraisals due to program activities explicitly designed to change negative appraisals (p_1). Second, there is an indirect effect of the program on negative appraisals through perceived social skills (p_2 and p_8). To determine program effects on negative appraisals, I need to take both sources into account. This is why I included line 29 in the Mplus syntax in Table 11.1. The relevant information for evaluating the program effect occurs in the output section `TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS`, subsection `Effects from TREAT to NEGAPP2`. Here is the output:

```
                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.    P-Value

Effects from TREAT to NEGAPP2

  Total              -1.132      0.053    -21.220      0.000
  Total indirect     -0.533      0.059     -8.981      0.000

  Specific indirect 1
    NEGAPP2
    PSKILLS2
    TREAT            -0.533      0.059     -8.981      0.000

  Direct
    NEGAPP2
    TREAT            -0.598      0.069     -8.712      0.000
```

The first line reports the total effect of the program on negative cognitive appraisals. The program lowered mean appraisals by -1.13 ±0.11 disagree-agree units relative to the control group ($z = 21.22$, $p < 0.05$), holding constant the relevant covariates in the equations. The second line of output reports how much of the effect of the treatment on negative cognitive appraisals is due to its impact through other mediators, in this case `PSKILLS2`, rather than directly. The reported path coefficient is -0.53 ±0.12, $z = 8.98$, $p < 0.05$. On the next three lines under `Specific indirect 1`, the relevant indirect effects are listed. The causal chain is read from the bottom up, from `TREAT` to `PSKILLS2` to `NEGAPP2`. By and of itself, the treatment reduces negative appraisals by -0.60 units (see the section `Direct`). Indirectly, the treatment reduces negative appraisals by -.53 units. These effects total to -1.13 units.

The meaningfulness standard for T→NEGAPP2 was ≤-0.75. The coefficient was more negative than this value but I need to take into account sampling error when making

my conclusion. The 95% confidence interval for T→NEGAPP2 was -1.24 to -1.03. The upper limit of the interval is less than -0.75, suggesting I can be reasonably confident the program had a meaningful effect on negative appraisals.

Because of the presence of both direct and indirect (through PSKILLS2) program effects on NEGAPP2, it is not straightforward to calculate the probability of exceptions to the rule and other standardized effect size indices. I discuss the relevant issues and the needed Mplus syntax to do so in a separate document on effects sizes for this chapter on the resources tab of my webpage. Using the methods described there, I found the probability of exceptions was 0.08. The general rule is that people who participate in the program tend to have lower negative cognitive appraisals than people who do not participate in the program. The level of exceptions to this rule is about 8%: For random draws of pairs of people, one from the treatment group and one from the control group, about 8% of the time, the person from the control group will have *lower* negative cognitive appraisals than the person from the treatment group, holding constant the relevant covariates. For the other standardized indices, see the standardized effect size document for the current chapter on my web page.

Effect of Program on External Locus of Control

Per Figure 11.2 the program affects external locus of control via two sources, (1) the direct effect of the treatment on it due to program activities explicitly designed to change external locus of control (p_3), and (2) the indirect effect of the program on external locus of control through perceived social skills (p_2 and p_9). I again need to consider both sources, hence, line 30 in the Mplus syntax in Table 11.1. Here is the relevant output:

```
                                                    Two-Tailed
                    Estimate       S.E.     Est./S.E.   P-Value

Effects from TREAT to EXTERN2

  Total              -0.369      0.050     -7.349      0.000
  Total indirect     -0.393      0.061     -6.422      0.000


  Specific indirect 1
    EXTERN2
    PSKILLS2
    TREAT            -0.393      0.061     -6.422      0.000

  Direct
    EXTERN2
    TREAT             0.024      0.078      0.310      0.757
```

The output has the same format as that for negative appraisals. The effect of the treatment is to lower external locus of control by -0.37 ±0.10 units on its -3 to +3 disagree-agree metric ($z = 7.35$, $p < 0.05$). Further inspection reveals that only one of the two sources of this program effect was statistically significant, namely the indirect effect of the treatment on external locus of control through perceived social skills. For the direct effect that was due to program activities aimed at external locus of control, the result was statistically non-significant (coefficient = 0.024 ±0.15, $z = 0.31$, $p < 0.76$). This suggests the program activities aimed at reducing external locus of control likely need to be revisited by the program designers. As you will see later, additional analyses suggest it is best to drop this program facet entirely.

In sum, the program produced meaningful change in the perceived social skills and the negative cognitive appraisals mediators, but not the external locus of control mediator. I place these results in broader context below.

Results of the Analysis: Effects of Mediators on the Outcome

The next question is whether each of the targeted mediators are relevant to the outcome. Traditionally, this question focuses on the magnitude and statistical significance of the covariate-adjusted path coefficients linking a mediator to the outcome. However, in the present example, the task is complicated because causal relationships exist among some mediators. After defining the meaningfulness standards for this facet of the RET, I consider the two mediators that do not causally affect another mediator, negative appraisals and external locus of control as interpretation for them is straightforward. I then turn to the perceived social skills mediator, which affects the two other mediators.

Meaningfulness Standards for Mediator Effects on the Outcome

I again engaged focus groups with staff and clients to define a meaningfulness standard for the effects of the mediators on the outcome. For each mediator, it was 0.30; negative 0.30 for PSS2 and positive 0.30 for NCA2 and EXT2.

Effect of Negative Cognitive Appraisals on Social Phobia

The output section relevant to the NEGAPP2→LSP3 link is the section called `MODEL RESULTS`. It appears in Table 11.2.

Table 11.2: Mplus Output for Mediator Effects on Social Phobia

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
LSP3 ON				
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
LSP1	0.347	0.072	4.835	0.000

The last three variables listed are the covariates, which I ignore because they are not part of my narrative. For negative cognitive appraisals, the covariate-adjusted path coefficient (p_4 in Figure 11.2) was 0.39 ±0.19, z = 4.10, p < 0.05. The coefficient indicates that for every unit increase on the -3 to +3 disagree-agree metric for negative appraisals, the social phobia mean (based on the clinician rating metric ranging from 0 to 5) is predicted to increase by 0.39 units, holding constant the covariates and other mediators. Stated in the opposite, if the program decreases negative appraisals by one unit, the mean social phobia should decrease by -0.39 units. A two-unit decrease in negative appraisals should lower the mean social phobia by (2)(-0.39) = -0.78 units; a three-unit decrease should lower the mean social phobia by (3)(-0.39) = -1.17 units.

The effect size standard for the coefficient was ≥0.30. The observed coefficient was larger than this standard but I need to take sampling error into account for purposes of making a conclusion. The 95% confidence interval for the path coefficient linking NEGAPP2 to social phobia was 0.20 to 0.58. The interval for the coefficient is not fully contained in the latitude of meaningfulness; it overlaps with the latitude of effect ambiguity. This means that I cannot confidently conclude that the coefficient exceeds the effect size standard given the amount of sampling error that is operating. I can confidently conclude that the effect is non-zero (because it is statistically significant), but I cannot confidently conclude it is meaningful.

In terms of standardized effect sizes, I used the program for the probability of exceptions to the rule on my webpage to gain perspectives on exceptions to the rule for this mediator. The generalized rule is that "people who are higher on negative cognitive appraisals tend also to have higher social phobia." The probability of exceptions to the rule, P_E, was 0.43. This means that for 43% of the cases, a person who is higher than the mean on negative cognitive appraisals is lower than the mean on social phobia, holding constant all the other predictors in the equation, namely the other mediators, baseline

social phobia, the treatment condition someone is in, biological sex, and a history of parental hypercriticism. The unique explained variance (squared semi-part correlation) in latent social phobia accounted for by negative cognitive appraisals was 0.02. See the resource tab on my web page and Chapter 10 for how I calculated these values.

Effect of External Locus of Control on Social Phobia

Next, I consider the mediator external locus of control. From Table 11.2, the path coefficient from external locus of control to social phobia was -.002 ±0.18, z = -0.02, p < 0.99. The path coefficient is statistically non-significant and is virtually zero. The effect size standard for the coefficient is ≥0.30. The coefficient clearly fails to meet this standard. The 95% confidence interval for the coefficient was -0.18 to 0.18. The confidence limits are completely outside the latitude of meaningfulness, indicating the observed coefficient is not meaningful.

Effect of Perceived Social Skills on Social Phobia

The analysis of the final mediator, perceived social skills, is complicated relative to the other two mediators because in addition to a direct effect on the outcome, it also can affect the outcome through its impact on the other two mediators. It is for this reason I included line 28 in the Mplus syntax in Table 11.1. The relevant information on the output for evaluating this mediator occurs in the section `TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS` and the subsection `Effects from PSKILLS2 to LSP3`. Here is the initial part of the output:

```
Effects from PSKILLS2 to LSP3

                                                    Two-Tailed
                  Estimate       S.E.    Est./S.E.  P-Value

Total              -0.883      0.084    -10.459      0.000
Total indirect     -0.177      0.056     -3.174      0.002
```

Underneath `Estimate` in the first line is the coefficient reflecting the effect of post-program perceived social skills on the latent social phobia at follow-up, allowing for all the different ways perceived social skills exerts that influence. The coefficient was -0.88 ±0.17, z = 10.46, p < 0.05. For every one unit increase on the averaged disagree-agree scale for perceived social skills, the latent variable social phobia mean is predicted to decrease by -0.88 units, holding constant the relevant covariates in the equation. A two unit increase in perceived social skills should lower the mean by (2)(-.88) = -1.76 units; a

three unit increase should lower the mean by (3)(-.88) = -2.64 units.

The second line of the output (labeled TOTAL INDIRECT) reports how much of this effect of perceived social skills on social phobia is due to its impact on the other mediators. The reported path coefficient is -0.18 ±0.11, z = 3.17, p < 0.05. For every one unit increase on the metric for perceived social skills, the mean social phobia is predicted to decrease by -0.18 units *through the other mediators*. As with the total effect, because the effect of perceived social skills on social phobia is a complex function of multiple path coefficients, it probably is best to use percentile bootstrapping to evaluate it. When I did so, the results were almost identical to those of the MLR approach.

The effect size standard for the coefficient was ≤-0.30. The sample coefficient of -0.88 is less than this standard. The 95% confidence interval for the coefficient linking PSKILLS2 to social phobia was -1.05 to -0.72. Because even the upper limit of the interval is less than the standard, I conclude that the PSKILLS2→LSP3 link is meaningful even given the operative sampling error.

In terms of standardized effect sizes, the calculation of the probability of exceptions and unique explained variance is complicated by the multiple ways that perceived social skills influences social phobia. I discuss the underlying issues and relevant Mplus syntax to estimate the statistics in the document for effect sizes for this chapter on the resources tab of my webpage. The probability of exceptions to the rule was 0.34.

In sum, based on the unstandardized coefficients, both negative cognitive appraisals and perceived social skills yielded statistically significant path coefficients relative to their effects on social phobia; the data also suggest meaningful effect sizes for perceived social skills. The path coefficient for external locus of control was not statistically significant.

Results of the Analysis: Unmeasured Mediators

There is one additional substantive point to address. In Table 11.2, the coefficient for the path from the treatment condition to LSP3 is -0.49 ±0.27 (z = 3.58, p < 0.05). This is the estimated effect of the program on social phobia independent of the three mediators directly targeted by the program. The mean latent social phobia at follow-up is -0.49 units lower for the treatment group than the control group *due to these unmeasured and unspecified mediators*. The meaningfulness of this effect is determined by the minimal meaningful change for social phobia (which is -1.0) times the fraction of the effect on social phobia that I believe this path should account for, per my discussion in Chapter 10. I might use a fraction of 0.25 given four predictors (less the covariates), which means the meaningfulness standard is (-1.0)(0.25) = -0.25. The coefficient of -0.49 exceeds -0.25 so it appears the treatment direct effect is meaningful. However, the 95% confidence

interval was -0.76 to -0.22, which overlaps the latitude of effect ambiguity. Taking sampling error into account, I can't confidently conclude there is a meaningful independent direct effect of the treatment on the outcome, although I can conclude that the effect is non-zero.

Summary of RET Results

At this point, I can piece together a summary of the FISEM-based RET analyses. The first issue I addressed was the overall impact of the program on social phobia. Using the metric of the clinician rating scale, Figure 11.4 shows the posttest mean for the control group on `CR3` as a reference point (dark arrow) and the estimated posttest mean for the treatment group relative to this reference point (the red arrow) -1.76 units to the left. The overall degree of change in social phobia due to the program is both statistically significant ($p < 0.05$) and meaningful, based on the meaningfulness standard of -1.0 set by the research team and staff. Patients, on average, moved from being clinically social phobic that is moderately disabling to being between mildly social phobic and moderately social phobic that is only somewhat disabling. To be sure, there is room for improvement; we would like to displace the red arrow even further to the left. However, I can confidently conclude the program is having a meaningful effect.

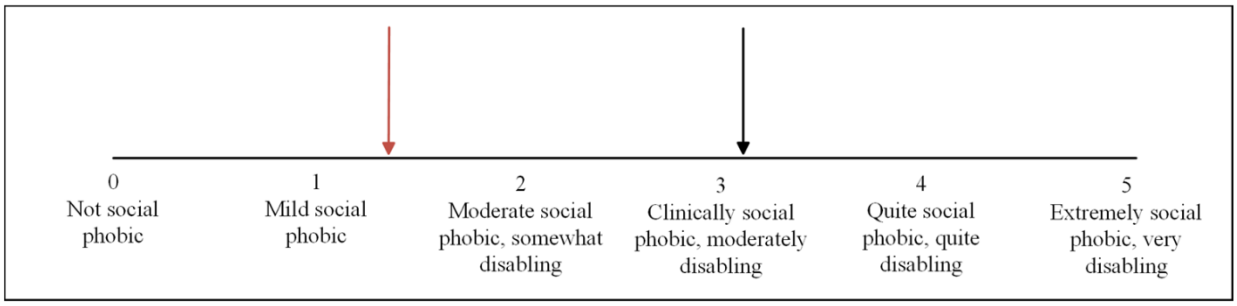

FIGURE 11.4. Line graph of total effect

The program sought to bring about change in three mediators/mechanisms, (1) negative cognitive appraisals, (2) perceived social skills, and (3) social-phobia based external locus of control, on the assumption that each is relevant to social phobia. A key question is whether these assumptions are viable. The data suggest a mixed picture. Figure 11.5 presents the -3 to + 3 disagree-agree metric for each mediator, which I make use of in my characterizations below. Keep in mind that this metric is an average across

multiple items, so changes in it reflect changes across the items considered as a totality.

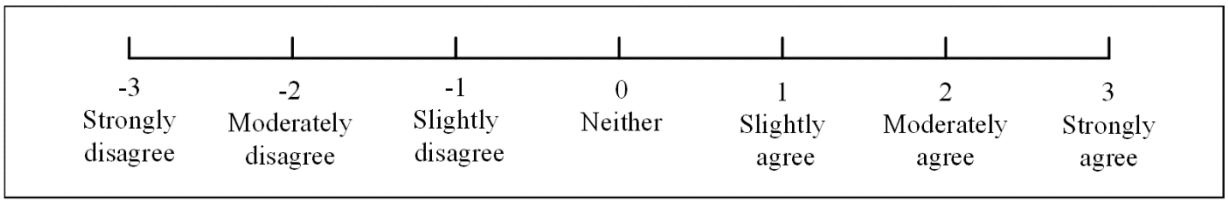

FIGURE 11.5. Mediator metric

For negative cognitive appraisals, I found that for every one unit the program is able to decrease negative cognitive appraisals on the above metric, the mean of social phobia decreases by about 0.39 units on the social phobia outcome metric, holding constant the other mediators and relevant covariates. This result was statistically significant and judged to represent a non-zero effect on social phobia. However, when taking sampling error into account, I could not confidently conclude in favor of a meaningful effect.

For the perceived social skills mediator, the coefficient associated with it exceeded its effect size standard even after sampling error was taken into account. I found that for every one unit the program is able to increase perceived social skills, the mean of social phobia decreases about 0.88 units on the clinician rating metric.

For external locus of control, the path coefficient linking it to social phobia was functionally zero. Everything points to the conclusion that efforts to bring about change in external locus of control for purposes of reducing social phobia should probably be abandoned, with resources focused elsewhere.

Finally, unmeasured and unspecified mediators of program effects were operating as reflected by the direct effect of the treatment condition on social phobia, holding constant the explicitly measured mediators and relevant covariates. The coefficient for the direct effect was -0.49, or about half a unit on the clinician rated social phobia metric. Taking into account the fraction of the overall effect I want the unmeasured mediators to account for, the coefficient exceeded its effect size standard. However, its 95% confidence interval overlapped with the latitude of effect ambiguity, indicating I could not confidently declare its meaningfulness. My conclusion is that the research team and staff should examine the program activities more closely to discern what mechanisms are driving the unmeasured mediators effect. Perhaps doing so will allow us to strengthen the program by adding activities that bring about even more change in them.

The final question I addressed was the extent to which the program was successful in bringing about meaningful change in the target mediators. The degree of change in the

external locus of control mediator is moot because, as noted above, the mediator appears to be of marginal relevance to social phobia for this particular population. For negative cognitive appraisals, the control group posttest mean disagree-agree response was 0.98, which represents an anchor point near "slightly agree." Using this as a reference point, the posttest mean for the intervention group was -1.13 disagree-agree units lower, or a value just below zero, which has an anchor of "neither agree nor disagree." The change was deemed meaningful, but clearly, there is room for improvement by moving the mean closer to a value of, say, -2 or -3.

The dynamic for perceived social skills was similar, except the program sought to increase rather than decrease the value of this mediator. The control group posttest mean disagree-agree response was -0.98, which represents an anchor near "slightly disagree." Using this as a reference point, the posttest mean for the program group was 1.17 disagree-agree units higher than this, or a value just above zero, which has an anchor of "neither agree nor disagree." The amount of change was deemed meaningful, but clearly, there is room for improvement by moving the mean closer to a value of 2 or 3.

The above results are informative and provide a rich framework for discussing with program staff how the program is doing and how to improve the program.

Traditional Mediation Analysis

Note that I have addressed the three major questions for program evaluation without performing classic tests of mediation as described in the mediation literature. Traditional mediation tests ask whether a given mediator, M, can account for some of the effects of a distal variable, T, on an outcome, Y. These tests traditionally evaluate the product of the causal coefficient reflecting the effect of T on M multiplied by the causal coefficient reflecting the effect of M on Y for each mediator. In my opinion, such omnibus tests in the context of RETs are not as helpful as analyzing the individual links in the mediational chain per the methods I have outlined. The omnibus tests provide little information gain beyond what I have already garnered with the tests of individual links. The omnibus test tells me if at least one of the links in a given mediational chain is "broken." However, it does not tell me which link is broken. A focus on the individual links of the chain pinpoints where the problem (broken link) is so that I can then decide if the link is fixable. If the problem is that the treatment does not meaningfully affect a given mediator, can program staff and/or we as scientists figure out how to change the treatment so that it does affect the mediator? Is it even worth trying to do so if we also learn that the mediator is irrelevant to the outcome? Might it be possible to alter the program to strengthen the causal coefficient linking a mediator to the outcome, per my discussion in Chapter 2? I find I can provide useful advice to program staff and

management by analyzing and juxtaposing the individual links of mediational chains, with only marginal information gain added by omnibus tests of mediation. By relegating omnibus tests to the substantive backyard, many of the statistical challenges of mediation analysis go with them, as will be apparent in future chapters. This is not to say scenarios do not exist in the social sciences where omnibus mediational tests are of interest. However, for purposes of program evaluation and providing feedback to program developers and administrators, a focus on the three facets of evaluation (does the program meaningfully affect the outcome; are the target mediators, in fact, relevant; does the program meaningfully affect the targeted mediators) is primary and illuminating.

Some scientists argue that omnibus indices of mediation are useful for identifying the most important mediators among a set of mediators. I disagree and do not think omnibus tests are the best way to approach such a question in the context of program evaluation. This is because omnibus tests confound (a) the effect of the treatment on each mediator with (b) the effect of each mediator on the outcome. I prefer to focus on the causal coefficients that link each mediator to the outcome to determine their relative importance; then I isolate the program effects on the most important, highest priority mediators. These two pieces of information provide more nuanced information than the single piece of information that comes with omnibus tests.

Finally, the focus on each individual link of a given mediational chain allows us to test and make statements about the statistical significance of the overall omnibus effect vis-à-vis the joint significance test as discussed in Chapter 9. The only unique information we obtain from the traditional omnibus index is its overall unstandardized effect size, which to me, does not add much beyond what I have already learned.

For the sake of completeness, I report the traditional omnibus tests in a document for this chapter on the resources tab of my webpage. This document also includes the analysis of interventional indirect effects mentioned in Chapter 9.

Sensitivity Analyses

In Chapter 5, I discussed the utility of conducting analyses to explore the sensitivity of one's conclusions to modeling assumptions. In the current example, I made several key assumptions about the absence of correlations between disturbances. One such assumption was a zero correlation between the negative cognitive appraisals disturbance term (d_1) and the latent social phobia disturbance term (d_4) in Figure 11.2. If the assumption is incorrect and the two disturbances are non-trivially correlated, then ignoring the correlation can bias the causal coefficient between negative cognitive appraisals and social phobia. My own belief is that the measured covariates I controlled for are sufficient to reduce the correlation between the disturbances to non-consequential

levels, but it would be reassuring if I conduct sensitivity analyses to determine how strong the correlation between the disturbances would have to be in order to meaningfully bias the target causal coefficient. It is possible to do such analyses in Mplus using FISEM. I provide a document *Sensitivity Tests in Full Information SEMs* on the resources tab of my webpage for the current chapter that walks you through the process in general and for the social phobia example in particular.

Competing Models

In Chapter 7, I discussed the concept of comparing one's model with viable competing models that are conceptually plausible. In the present case, no strong alternative conceptual models suggest themselves with the possible exception of the causal arrows from PSKILLS2 → NEGAPP2 and PSKILLS2 → EXTERN2 being reversed in direction. That is, instead of perceived social skills impacting negative cognitive appraisals, the reverse might be true and similarly so for EXTERN2. Possibly even more likely is that there are reciprocal causal relationships between the variables such that, for example, PSKILLS2 → NEGAPP2 *and* NEGAPP2 → PSKILLS2. It is good practice to explore such competing models when possible or, at the very least, to recognize their plausibility as competing models in one's write-up. I present such analyses for the current case in a document on competing models on the resources tab of my webpage for this chapter. It turns out there was no strong evidence for models with reverse causality.

Measurement Error for Single Indicators

I was able to adjust for measurement error in my analyses for social phobia by using multiple, interchangeable indicators of the social phobia construct. However, I did not adjust for measurement error for the single indicator measures in the model, including negative cognitive appraisals, perceived social skills, and external locus of control as measured at baseline and posttreatment. In Chapter 3, I noted strategies in SEM for adjusting for measurement error for constructs measured by single indicators and provided a primer that explains how to do so on the resources tab of my website for Chapter 3. I applied the strategy to the social phobia example in the current chapter and present the relevant Mplus syntax and results on the resources tab for Chapter 11. The fundamental conclusions of the RET analyses were not affected when I did so.

Concluding Comments for Traditional FISEM Analyses

The FISEM analyses outlined in this section move well beyond the typical mediation analyses you will encounter in the program evaluation literature. I formulated an explicit

conceptual logic model for the program I was evaluating, articulating the key mechanisms through which the program was assumed to impact the program outcome. The conceptual logic model took the form of an influence diagram. In the model, I was careful to articulate plausible causal relationships among the mediators based on theory and past research. I also considered confounds and assumptions underlying sequential ignorability for purposes of choosing covariates to include in the model.

With the model in hand, I conducted preliminary analyses to ensure that the working assumptions I was making when modeling were viable. I performed checks for ill-behaved outliers and leverages that might distort my conclusions and ensured my assumptions about linearity were reasonable. I obtained multiple, interchangeable indicators of my primary outcome variable, although practical constraints prevented me from doing so for my mediators. However, the measures of the mediators had a proven track record and I was confident they had high levels of reliability and validity. I used a robust estimation method so that assumptions of normality and variance heterogeneity were minimized.

I fit the model to the data and ultimately evaluated it using the weight-of-evidence perspective outlined in Chapter 7. I considered (a) the prior evidence for the model, (2) the fit of the model as reflected by diverse global fit indices, (3) the fit of the model as reflected by localized fit indices, (4) whether the predicted paths in the model were statistically significant and non-trivial, (5) whether the model made substantive sense, and (6) ruling out competing models. The original model I formulated was not supported by the data in the sense that the external locus of control mediator did not behave in ways I thought it would.

I addressed three questions in my program evaluation, (1) did the program have an effect on the targeted outcome, (2) did the program affect the mediators that its logic model identified as key to outcome change, and (3) were the targeted mediators, in fact, related to the outcome. I did not rely solely on statistical significance to address these matters. I formulated effect standards for meaningfulness and evaluated results relative to those standards, taking into account sampling error. I made note of margins of error (confidence intervals) for model parameters and incorporated them into my decision making.

Finally, I performed supplemental analyses (sensitivity analyses, evaluation of competing models, measurement error adjustments for single indicator constructs) to further increase my confidence in my conclusions. In future chapters, I discuss additional supplemental analyses (e.g., based on statistical power) that I routinely apply.

BAYESIAN SEM

In this section, I show you how to apply Bayesian SEM to the social phobia example. I assume you have read the section on Bayesian SEM in Chapter 8. If not, do so. Bayesian SEM is a form of FISEM. I describe how to apply it in its simplest form with non-informative priors defined by Mplus defaults. I develop the case of informative priors in Chapter X. My primary goals are to compare the results I obtain for Bayesian SEM for the current example with those that I presented above for traditional FISEM and to give you an introduction to Bayesian modeling with RETs. For the Bayesian application as well as the other analytic strategies I consider in the remainder of this chapter, I will not revisit the application of meaningfulness standards as doing so follows directly from the methods already discussed.

To conduct a Bayesian SEM, I use the same syntax as in Table 11.1 for traditional FISEM but with a few modifications. First, I change line 12 from `ESTIMATOR = MLR` to

```
ESTIMATOR = BAYES; BITERATIONS=100000 (50000); BCONVERGENCE =.01;
```

The `BAYES` specification requests the Bayes analysis. The two other options override technical defaults used by Mplus and which I recommend you use more generally (see McNeish, 2016, for the rationale). I remove `SAMP` and `MOD(ALL 4)` from the output line because these options are not allowed in Mplus with Bayesian models. I add the option `TECH8` to the output line, which then produces the PSR and Kolmogorov–Smirnov statistics for convergence. I also change the `CINTERVAL` option on the `OUTPUT` line to read `CINTERVAL(HPD)` to obtain asymmetric credible intervals. Finally, after line 31, I add a new line to the program to generate relevant plots, as discussed in Chapter 8.

```
PLOT: TYPE = PLOT2;
```

In the Mplus output, I first check the potential scale reduction (PSR) statistics to ensure the iterative process converged, per my discussion in Chapter 8. PSR values less than 1.1 suggest convergence. The relevant output is in the `TECHNICAL 8` output section. Here is a portion of it:

```
TECHNICAL 8 OUTPUT FOR BAYES ESTIMATION

                    POTENTIAL        PARAMETER WITH
  ITERATION     SCALE REDUCTION       HIGHEST PSR
  100               1.617                4
  200               1.479                1
  300               2.020                4
   .          .          .
```

```
      .          .          .
      .          .          .
   49600          1.003          5
   49700          1.003          5
   49800          1.003          5
   49900          1.004          5
   50000          1.004          5
```

I show the first few listings in the output and then the final listings after the … notation. Of primary interest are the iterations at the end, where the highest PSR should stabilize and be close to 1.0. The pattern of PSRs observed here suggests convergence. No problematic KS tests were printed out by Mplus.

The Bayesian analog of the chi square test of fit is the **posterior predictive p-value**. A questionable model is suggested by a p value < 0.05. Here is the relevant output:

```
MODEL FIT INFORMATION

          Posterior Predictive P-Value              0.606
```

The p value was 0.606, which is consistent with a reasonable fitting model. Mplus also reports a 95% confidence interval for the difference between observed and replicated chi-square values (see Chapter 8). Here is the output for it:

```
Bayesian Posterior Predictive Checking using Chi-Square

          95% Confidence Interval for the Difference Between
          the Observed and the Replicated Chi-Square Values

                              -46.147            36.289
```

The confidence interval should contain zero and the value of zero should fall near the middle of the interval. Reasonable fit is affirmed. Here is the output for additional fit indices that are more familiar to us:

```
RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.000
          90 Percent C.I.                    0.000  0.030
          Probability RMSEA <= .05           0.999

CFI/TLI

          CFI                                1.000
          90 Percent C.I.                    0.995  1.000
```

The RMSEA, the p value for close fit, and the CFI all point to satisfactory model fit.

Here are the predicted correlations for the observed variables from the output:

```
           Correlations
 CR1            1.000
 SPAI1          0.794         1.000
 SPIN1          0.819         0.796         1.000
 CR3            0.165         0.160         0.165         1.000
 SPAI3          0.162         0.157         0.162         0.848         1.000
 SPIN3          0.167         0.162         0.167         0.874         0.860
 NEGAPP2        0.034         0.033         0.034         0.655         0.644
 PSKILLS2      -0.023        -0.023        -0.023        -0.708        -0.696
 EXTERN2        0.060         0.058         0.060         0.347         0.342
 NEGAPP1        0.149         0.145         0.150         0.099         0.098
 PSKILLS1      -0.180        -0.175        -0.180        -0.169        -0.166
 EXTERN1        0.150         0.146         0.150         0.067         0.066
 HYPER         -0.001        -0.001        -0.001        -0.003        -0.003
 SEX           -0.082        -0.080        -0.082        -0.047        -0.046
 TREAT          0.042         0.041         0.042        -0.638        -0.627

           Correlations
              SPIN3         NEGAPP2       PSKILLS2      EXTERN2       NEGAPP1
              ________      ________      ________      ________      ________
 SPIN3          1.000
 NEGAPP2        0.664         1.000
 PSKILLS2      -0.717        -0.749         1.000
 EXTERN2        0.352         0.386        -0.492         1.000
 NEGAPP1        0.101         0.285        -0.054         0.109         1.000
 PSKILLS1      -0.171        -0.175         0.279        -0.206        -0.204
 EXTERN1        0.068         0.107        -0.072         0.377         0.216
 HYPER         -0.003         0.121        -0.068         0.197         0.231
 SEX           -0.047        -0.052         0.032        -0.002        -0.041
 TREAT         -0.646        -0.717         0.759        -0.355         0.006

           Correlations
              PSKILLS1      EXTERN1       HYPER         SEX           TREAT
              ________      ________      ________      ________      ________
 PSKILLS1       1.000
 EXTERN1       -0.178         1.000
 HYPER         -0.220         0.326         1.000
 SEX           -0.050        -0.003         0.026         1.000
 TREAT         -0.015        -0.031        -0.022         0.023         1.000
```

and here are the observed correlations, which should be similar to the above:[4]

[4] I used the Mplus analysis type BASIC to obtain these correlations in a separate run.

Correlations

	CR1	SPAI1	SPIN1	CR3	SPAI3
CR1	1.000				
SPAI1	0.789	1.000			
SPIN1	0.816	0.790	1.000		
CR3	0.183	0.170	0.144	1.000	
SPAI3	0.164	0.158	0.134	0.845	1.000
SPIN3	0.196	0.172	0.174	0.870	0.859
NEGAPP2	0.090	0.026	0.019	0.653	0.630
PSKILLS2	-0.072	-0.022	0.002	-0.715	-0.696
EXTERN2	0.113	0.081	0.082	0.341	0.322
NEGAPP1	0.149	0.150	0.146	0.087	0.084
PSKILLS1	-0.207	-0.182	-0.149	-0.189	-0.153
EXTERN1	0.129	0.167	0.151	0.148	0.090
HYPER	-0.005	-0.010	0.008	-0.005	-0.011
SEX	-0.068	-0.070	-0.102	-0.033	-0.068
TREAT	0.022	0.049	0.056	-0.633	-0.618

Correlations

	SPIN3	NEGAPP2	PSKILLS2	EXTERN2	NEGAPP1
SPIN3	1.000				
NEGAPP2	0.667	1.000			
PSKILLS2	-0.704	-0.738	1.000		
EXTERN2	0.369	0.367	-0.489	1.000	
NEGAPP1	0.085	0.264	-0.007	0.051	1.000
PSKILLS1	-0.157	-0.193	0.278	-0.194	-0.205
EXTERN1	0.110	0.173	-0.071	0.375	0.217
HYPER	0.003	0.121	-0.068	0.196	0.231
SEX	-0.043	-0.052	0.032	-0.002	-0.041
TREAT	-0.651	-0.718	0.756	-0.352	0.006

Correlations

	PSKILLS1	EXTERN1	HYPER	SEX	TREAT
PSKILLS1	1.000				
EXTERN1	-0.178	1.000			
HYPER	-0.221	0.326	1.000		
SEX	-0.050	-0.003	0.026	1.000	
TREAT	-0.015	-0.032	-0.022	0.022	1.000

Table 11.3 presents the path coefficients for the original MLR analysis and the corresponding ones for the Bayes model. The analyses produced comparable results, as often happens for uninformative priors. The Bayes output does not provide z or p values. If the credible interval does not contain zero, the coefficient is said to be statistically significant, $p < 0.05$, although null hypothesis testing is not core to Bayesian philosophy.

Table 11.3: MLR and Bayesian Parameter Estimates

	FISEM MLR		*FISEM Bayes*	
Parameter	Coef	95% CI	Coef	95% CI
T → NCA2 (p_1)	-0.60	-0.73 to -0.46	-0.60	-0.75 to -0.45
PSS2 → NCA2 (p_8)	-0.46	-0.55 to -0.36	-0.46	-0.55 to -0.36
T → PSS2 (p_2)	1.17	1.08 to 1.27	1.17	1.07 to 1.27
T → ELC2 (p_3)	0.02	-0.13 to -0.18	0.02	-0.12 to 0.17
PSS2 → ELC2 (p_9)	-0.34	-0.43 to 0.24	-0.34	-0.43 to -0.24
T → SP3 (p_7)	-0.49	-0.76 to -0.21	-0.49	-0.78 to -0.20
NCA2 → SP3 (p_4)	0.39	0.20 to 0.58	0.39	0.21 to 0.57
PSS2 → SP3 (p_5)	-0.71	-0.90 to -0.51	-0.71	-0.91 to -0.50
ELC2 → SP3 (p_6)	0.00	-0.18 to 0.18	0.00	-0.99 to 0.19

Notes: Coef=coefficient; CI=confidence/credible interval; SP=social phobia latent variable; T=treatment group; NCA=negative cognitive appraisals; PSS=perceived social skills; ELC=external control

As described in Chapter 8, you can use the plot menu in Mplus to visualize a kernel density plot of the posterior distribution for any parameter. Figure 11.6 presents the plot for the path coefficient from the negative cognitive appraisal mediator to latent social phobia, namely p_4. (I moved the legend to the right; Mplus overlays it on top of the distribution). The information in the legend provides the mean, median, mode and standard deviation of the posterior distribution as well as the values of the 95% credible intervals. The posterior distribution in this case is bell shaped and centered around the indices of central tendency. The leftmost and rightmost vertical lines in the plot area per se reflect the lower and upper limits of the credible interval, respectively.

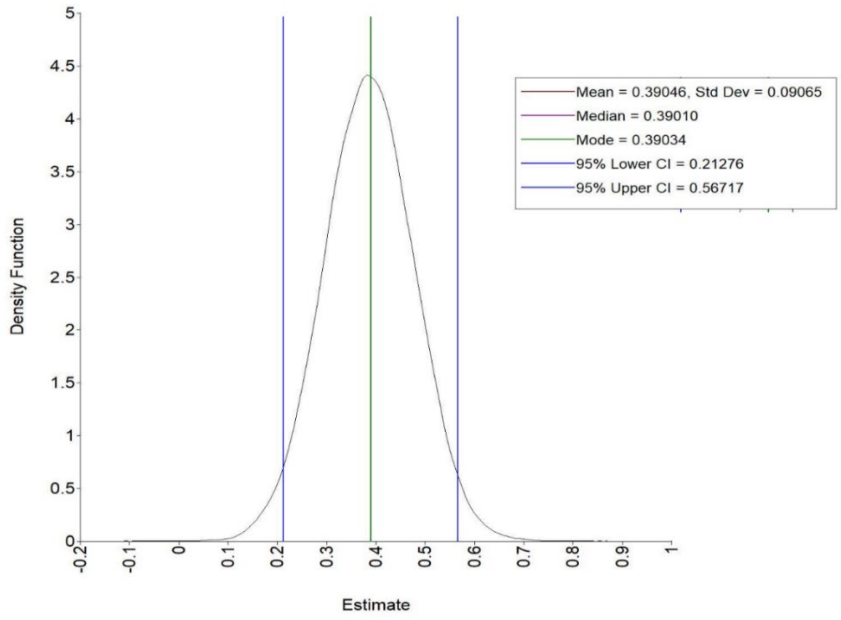

FIGURE 11.6. Example of posterior distribution

From these results, you can replicate all of the analyses from the traditional FISEM analysis to answer the three central questions for an RET (what is the total effect of the program on the outcome, what is the effect of the mediators on the outcome, what is the effect of the program on the mediators). I do not do so here in the interest of space. The Bayesian analysis also produces feedback on the measurement model from the example.

Effect size indices as discussed in Chapter 10 have received relatively little attention in Bayesian SEM. The approach I described for the analysis of unstandardized effect size analysis can be used but with credible intervals instead of confidence intervals. My approach to using latitudes is similar to a method known as the region of practical equivalence (ROPE) in the Bayesian literature (see Kruschke & Liddell, 2018). Jeffreys (1961) has advocated for the use of the Bayes factor statistic as a standardized index of effect size, but it is challenging to calculate for applications like the one in this chapter. For details, see the document for standardized effect size estimation on my webpage.

Some analysts prefer Bayesian modeling to more traditional modeling because it allows you to take into account prior information about the parameters (although I did not do that here), it approaches missing data in an elegant way (see Chapter XX), and the interpretation of credible intervals is more straightforward than confidence intervals.

LIMITED INFORMATION SEM

In this section, I apply limited information estimation approaches to the social phobia example. I have redrawn Figure 11.2. in Figure 11.7 to remove the latent variables because many of the LISEM analyses focus only on single indicator models. I use CR3 as the outcome variable because its metric is intuitive and meaningful to my client.

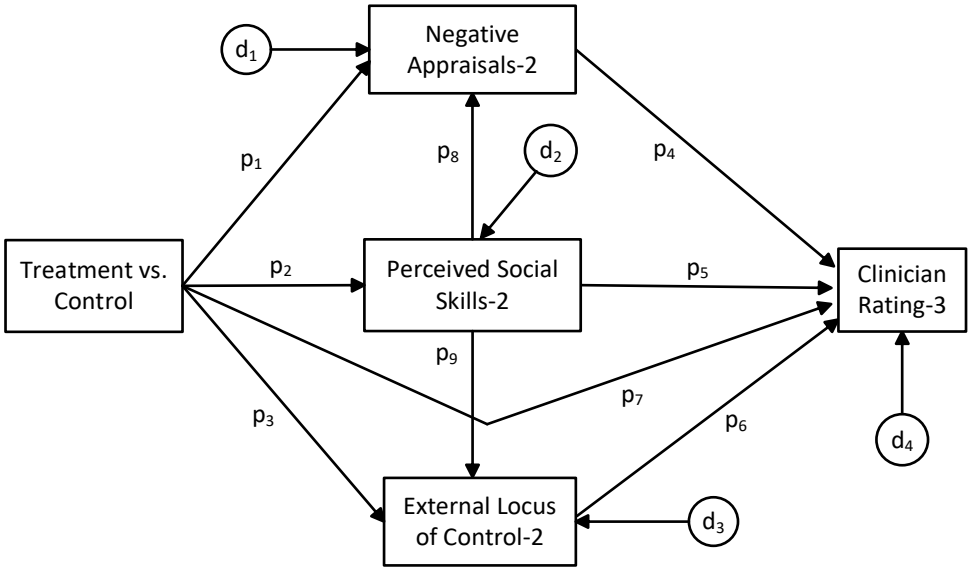

FIGURE 11.7. Social phobia example with single indicators

The equations implied by the influence diagram are:

$$NCA2 = a_1 + p_1\, T + p_8\, PSS2 + b_1\, BS1 + b_2\, PH1 + b_3\, NCA1 + d_1 \quad [11.11]$$

$$PSS2 = a_2 + p_2\, T + b_4\, BS1 + b_5\, PH1 + b_6\, PSS1 + d_2 \quad [11.12]$$

$$ELC2 = a_3 + p_3\, T + p_9\, PSS2 + b_7\, BS1 + b_8\, PH1 + b_9\, ELC1 + d_3 \quad [11.13]$$

$$CR3 = a_4 + p_7\, T + p_4\, NCA2 + p_5\, PSS2 + p_6\, ELC2 + b_{10}\, BS1 + b_{11}\, PH1 + b_{12}\, CR1 + d_4 \quad [11.14]$$

As with Figure 11.2, I label the path coefficients with *p*s and the covariates that are not part of the theoretical narrative with *b*s. For LISEM, I conduct separate regression analyses, one for each of the above equations. In this section, I first apply ordinary least squares regression LISEM to the social phobia data. I might use this approach if I am concerned about the sample size being too small to accommodate asymptotic theory. I

then explore the use of quantile regression, robust regression, Bollen's MIIV-SEM approach, and a LISEM version of Bayesian analysis. Quantile regression allows me to analyze outcome medians instead of means, perhaps to deal with outliers. I also can explore effects of the intervention at different segments of the outcome distribution. Robust regression allows me to deal with outliers and non-normality in ways that robust maximum likelihood can not. For Bollen's MIIV-SEM approach, I revert to Figure 11.2 and make use of the latent variables even though the approach is a form of LISEM. The use of only a single indicator of social phobia in the other models sacrifices the advantages of multiple indicators of SEM, but LISEM comes with advantages as well, as discussed in Chapter 8. By using the reference indicator represented by the clinician measure, I am able to directly compare results for the FISEM and LISEM. The clinician ratings had low levels of measurement error, which helps. The LISEM version of Bayes might be used if I feel the model I am testing is too complex given the sample size.

LISEM: Ordinary Least Squares Regression

For OLS regression, I used SPSS to conduct the regression analyses for the four equations. Table 11.4 presents the path coefficients for the equations from the original latent-variable FISEM analysis and the corresponding coefficients from the OLS analyses. There is close correspondence between the coefficients.

In this section, I first evaluate model fit and then address the three questions core to RETs, namely (1) is there an effect of the program on the outcome, (2) does the program affect the mediators it is intended to affect, and (3) do the mediators affect the outcome?

Table 11.4: Comparison of FISEM and LISEM

	Original FISEM		*OLS LISEM*	
Parameter	Coefficient	CR	Coefficient	CR
T → NCA2 (p_1)	-0.60 ±0.14	8.71*	-0.60 ±0.15	7.88*
PSS2 → NCA2 (p_8)	-0.46 ±0.10	9.52*	-0.46 ±0.10	9.22*
T → PSS2 (p_2)	1.17 ±0.10	23.64*	1.17 ±0.10	23.49*
T → ELC2 (p_3)	0.02 ±0.16	0.31	0.02 ±0.15	0.33
PSS2 → ELC2 (p_9)	-0.34 ±0.10	6.71*	-0.34 ±0.10	7.06*
T → SP3 (p_7)	-0.49 ±0.27	3.58*	-0.42 ±0.33	2.53*

NCA2 → SP3 (p_4)	0.39 ±0.19	4.10*	0.38 ±0.20	3.76*
PSS2 → SP3 (p_5)	-0.71 ±0.20	7.11*	-0.77 ±0.23	6.74*
ELC2 → SP3 (p_6)	0.00 ±0.18	0.02	-0.03 ±0.21	0.26

Notes: CR=critical ratio; SP=social phobia (latent variable for FISEM, clinician rating otherwise); T= treatment condition; NCA=negative appraisals; PSS=perceived social skills; ELC=external locus of control; * $p < 0.05$

Evaluation of Model Fit

I evaluated model fit using the strategy of independence tests discussed in Chapter 8. I used the program *graph theory* (DAGitty) on my website to identify the implied independence tests for the model in Figure 11.7. For example, one model-based independence assumption is that the association between negative cognitive appraisals at posttest and external locus of control at posttest should be zero if I hold constant (a) perceived social skills at posttest, (b) the treatment condition, (c) the baseline negative cognitive appraisals, (d) biological sex, and (e) the history of parental hypercriticism. I calculated this partial correlation and found it to equal -0.006 with a p value of 0.92, which is consistent with model predictions.[5] The graph theory program identified 23 independence assumptions after taking into account all the covariates and variables of substantive interest. Each assumption is evaluated to gain perspectives on model fit. Given the number of contrasts, you probably want to take into account chance associations using one of the strategies discussed in Chapter 6. Across the contrasts, the data for the social phobia example were model supportive. My preference is to work at the level of these localized tests rather than combining them into the omnibus C statistic described in Chapter 8. You can download the DAGitty code I used on the resources tab for Chapter 11 of my website.

Analysis of the Total Effect

As discussed in Chapter 8, there are two ways I can estimate the total effect using OLS-based LISEM. The simplest strategy, the **direct regression method**, is to regress `CR3` onto the treatment condition, `TREAT`, the baseline clinician rating, CR1, and the two covariates of biological sex and parental hypercriticism using standard OLS regression. The coefficient for `TREAT` is the estimated total effect of the program on social phobia. The coefficient was -1.73 ±0.22 ($t(328) = 15.41$, $p < 0.05$). This compares well to the

[5] Some LISEM frameworks do not embrace partial correlations per se, but one can still test for independence using carefully constructed regression equations because p values for partial regressed coefficients equal the p values for partial and semi-part correlations. In the present case, I would regress NCA2 onto EXTERN2 and the other covariates and evaluate the p value for the coefficient associated with EXTERN2.

estimate from the FISEM analysis where the result was -1.76 ±0.21 ($z = 16.86$, $p < 0.05$).

The second method, the **combined coefficient method**, uses model-based Monte Carlo confidence intervals per Chapter 8. The total effect of the program on social phobia is a function of path coefficients in the model combined via the following equation:

$$\text{Total Effect} = p_1p_4 + p_2p_5 + p_3p_6 + p_2p_8p_4 + p_2p_9p_6 + p_7$$

I used the program *Monte Carlo CIs* on my website, entering the above expression and the relevant asymptotic covariance matrix for the path coefficients on the right side of the equation as taken from OLS output (see the video for the program for details). The result was -1.74 ±0.24, $p < 0.05$.

For a standardized index of effect size, I used the program on my website for exceptions to the rule for the results yielded by the direct regression method. I found the value of P_E to be 0.11. The general rule is that people who participate in the program tend to have lower social phobia than people who do not participate in the program. The level of exceptions to this rule, expressed as a percent, is about 11%. For random draws of pairs of people, one from each group, about 11% of the time there will be exceptions to the rule, holding constant baseline social phobia, biological sex, and the person's history of parental hypercriticism. This maps well onto to the result for the FISEM analysis. For calculation of Cohen's *d* and the proportion of unique explained variance, see the document on standardized effect sizes on the resources tab of my webpage.

Analysis of Program Effects on Mediators

When other mediators do not impact the mediator of interest, the effect of the program on the mediator is the value of the path coefficient linking the treatment condition to the mediator in the relevant OLS equation. This was the case for perceived social skills, whose regression coefficient in Equation 11.13 was 1.17 ±0.10 ($z = 23.49$, $p < 0.05$; see Table 11.4). The coefficient in the FISEM analysis was 1.17 ±0.10 ($z = 23.64$, $p < 0.05$), which is comparable. The standardized effect size indices can be computed using the methods described in the document on standardized effect sizes on the resources tab of my website. For example, I used the program called *Prob of exceptions: Binary* on my website and found the probability of exception to be 0.03. The general rule is that people who participate in the intervention will score higher than people who do not participate in the intervention. The percent of exceptions to this rule is 3%, i.e., if I randomly select a person from the treatment group and a person from the control group, across multiple random draws, about 3% of the time the person in the control condition will have a higher perceived social skills score than the person in the treatment condition, holding constant the baseline perceived social skills, biological sex, and hypercritical parenting.

When a mediator is impacted by other mediators that also are impacted by the intervention, then this dynamic needs to be taken into account to estimate the full effect of the program on the mediator of interest after controlling for nuisance covariates. There are two ways that the multiple causes can be accommodated. The methods map onto the two strategies for estimating the total effect using OLS in the previous section. I illustrate both approaches using the negative cognitive appraisals mediator.

One strategy is to regress the mediator onto the program dummy variable plus the relevant covariates but omitting the mediator(s) that also determines the target mediator so as not to hold it constant. In this case, the regression equation omits perceived social skills at the posttest and is:

$$NCA2 = a + p\,T + b_1\,BS1 + b_2\,PH1 + b_3\,NCA1 + d$$

The coefficient for the treatment condition was -1.13 ±0.11 ($t(328) = 20.28$, $p < 0.05$). The estimate from the FISEM analysis was -1.13 ±0.11 ($z = 21.22$, $p < 0.05$), which is comparable. As before, I refer to this method as the **direct regression method**.

The second approach is to apply the Monte Carlo confidence interval method using the estimated coefficients for p_1, p_2, and p_8 in Table 11.6 in the expression based on path tracing (see the primer on the resources tab of my website):

$$\text{Effect of program on NCA2} = p_1 + (p_2)(p_8)$$

Using the *Monte Carlo CI* program on my website, the effect of the program on negative appraisals was estimated to be -1.13 ±0.13, $p < 0.05$), which comports well with the FISEM result. As before, I refer to this approach as the **combined coefficient method**.

For a standardized index of effect size for program effects on negative cognitive appraisals, I used the program on my webpage for exceptions to the rule coupled with the direct regression method. I found the value of P_E to be 0.06. For the calculation of Cohen's *d* and the proportion of unique explained variance, see the document on standardized effect sizes for this chapter on the resources tab of my webpage.

Applying the same logic to the external locus of control mediator, the value of the treatment condition coefficient was -0.37 ±0.10 ($t(328) = 7.15$, $p < 0.05$) for the direct regression method and for the Monte Carlo confidence interval method it was -0.38 ±0.10, $p < 0.05$). The estimate from the FISEM analysis was -0.37 ±0.10 ($z = 7.35$, $p < 0.05$). All three estimates are comparable. See the standardized effect size document on my website for the calculation of standardized effect size indices.

Analysis of Mediator Effects on Outcomes

When a mediator influences the outcome directly and not through other mediators, its estimated effect on the outcome is its coefficient in the equation in which it embedded. In the current example, this is the case for both negative cognitive appraisals (p_4 in Equation 11.14) and external locus of control (p_6 in Equation 11.14). The coefficient for negative cognitive appraisals was 0.38 ±0.20 ($t(324) = 3.76$, $p < 0.05$), which corresponds well with the result from the FISEM analysis whose coefficient was 0.39 ±0.19, $z = 4.10$, $p < 0.05$); see Table 11.4. The coefficient for external locus of control was -0.03 ±0.21 ($t(324) = 0.26$, ns), which corresponds well to the result from the FISEM analysis whose coefficient was 0.00 ±0.18, $z = 0.02$, ns); see Table 11.4.

For perceived social skills, there are three sources of its effects on social phobia, (1) the direct effect it has on social phobia (p_5), (2) the indirect effect it has on social phobia through negative cognitive appraisals (p_8 p_4), and (3) the indirect effect it has on social phobia through external locus of control (p_9 p_6). One can estimate the full effect of perceived social skills on social phobia using the two methods described earlier. The direct regression method regresses the outcome onto the mediator of interest (perceived social skills) plus the relevant covariates but omitting the mediator(s) through which it influences the outcome, in this case, negative cognitive appraisals and external locus of control. Using the notation from Equation 11.14, the regression equation is:

$$CR3 = a + p_7\ T + p\ PSS2 + b_{10}\ BS1 + b_{11}\ PH1 + b_{12}\ CR1 + d$$

The path coefficient for perceived social skills was -0.92 ±0.20 ($t(327) = 9.25$, $p < 0.05$), which compares favorably to the result for the FISEM analysis (coefficient = -0.88 ±0.17, $z = 10.46$, $p < 0.05$). The second method is the combined coefficient method coupled with Monte Carlo confidence intervals for the expression (Figure 11.13)

$$\text{Effect of PSS2 on CR3} = p_5 + (p_8)(p_4) + (p_9)(p_6)$$

The estimated effect of perceived social skills on social phobia for this approach was -0.89 ±0.19, $p < 0.05$, which compares favorably with the FISEM result.

For the standardized effect size indices for the mediators predicting social phobia, see the standardized effect size document on the resources tab of my web page.

Analysis of Unmeasured Mediators

Finally, the OLS-based regression for Equation 11.14 contains an estimate of the direct effect of the treatment condition on social phobia independent of the measured mediators, p_7. The direct effect was -0.42 ±0.33 ($t(324) = 2.53$), which compares favorably with the

result from the FISEM analysis, which was -0.49 ±0.27, z = 3.58.

In sum, I addressed the three core questions of an RET but instead of using FISEM, I used OLS-based LISEM. In the present case, the results for the two forms of analysis were comparable. For complex models with small sample sizes and where FISEM is not viable, it often is possible to use LISEM as an alternative analytic approach, a topic I discuss in more detail in Chapter X. I did not address here how to apply OLS-based LISEM to omnibus tests of mediation. I tend to rely on the joint significance test in such cases, but for more elaborate approaches, see the documents on the Resources tab on my webpage.

Profile Analyses

In this section, I provide an example of a statistical tool that can be used with OLS-based LISEM but that is more challenging to apply in FISEM, especially with latent variables.

I sometimes find it helpful to present multivariate profiles of hypothetical study participants that take the form of counterfactuals to give the program evaluation staff a better sense of program multivariate dynamics. On my website, I provide a program called *profile analysis* that uses the marginal effect framework discussed in Chapter 5 (Leeper, 2021). Consider Equation 11.14 for predicting the posttest clinician rating from the posttest mediators, the treatment condition, and the covariates, which I repeat here:

$$CR3 = a_4 + p_7\,T + p_4\,NCA2 + p_5\,PSS2 + p_6\,ELC2 + b_{10}\,BS1 + b_{11}\,PH1 + b_{12}\,CR1 + d_4$$

The OLS regression analysis yielded an equation with the following values:

$$CR3 = 1.226 + -0.415\,T + 0.380\,NCA2 + -0.771\,PSS2 + -0.028\,ELC2 + 0.017\,BS1 + -0.188\,PH1 + 0.251\,CR1 \qquad [11.15]$$

A substantial number of patients (about 45%) in the control group had posttest values close to 1.0 for both negative cognitive appraisals and external locus of control, which corresponds to ratings of “slightly agree” on the -3 to +3 metric of these scales. They also had a value near -1.0 for perceived social skills, which maps onto a response of “slightly disagree” on its -3 to +3 metric. I can ask the question of what the average posttest clinician rating would be if everyone in the sample was in the control group and they all had values of 1.0 on the posttest negative cognitive appraisals, 1.0 on external locus of control and -1.0 on perceived social skills. Using the *profile analysis* program in conjunction with Equation 11.15, the program first calculates a predicted CR3 score for each individual in the data set but where (a) everyone’s T score is set to 0 to mimic the case where they had not completed the intervention program and (b) everyone is assigned

a NCA2 score of 1.0, a PSS2 score of -1.0, and an ELC2 score of 1.0, while (c) weighting all other covariate scores for a given individual by the associated regression coefficient for the respective covariate in Equation 11.15. The average of the generated predicted scores is then computed to reflect the "typical" clinician rating one would expect for patients with this multivariate profile. It was 3.14 ±0.14. On the CR3 metric, this score is near the anchor "clinically social phobic, moderately disabling" and represents the scale anchor where a substantial number of untreated patients would be given they have this multivariate mediator profile.

Next, I contrasted this profile, which I call Profile 1, with a set of different counterfactually defined profiles where everyone participates in the program but now where each mediator takes on values that are more favorable to healthy adjustment by differing amounts. I set the value of T to 1 for everyone and their scores on the three mediators to differing values of NCA2, PSS2, and ELC2 but again allowing individuals' other covariate scores to be weighted by their associated regression coefficients. For ELC2, I know from the prior analyses that it is trivially related to CR3 and, furthermore, that the program generally failed to bring about change in it. I therefore keep ELC2 constant at a value of 1.00 for the additional profiles I explore. Here are some profiles I created to gain a sense of the multivariate dynamics:

	NCA2	PSS2	ELC2	TREAT	Mean CR3
Profile 1	1.00	-1.00	1.00	0	3.14 ±0.14
Profile 2	0.50	-0.50	1.00	1	2.15 ±0.21
Profile 3	0.25	-0.25	1.00	1	1.86 ±0.18
Profile 4	0.00	0.00	1.00	1	1.58 ±0.15

Profile 2 relative to Profile 1 shows a patient profile in which individuals participated in the program and showed a net scale improvement of half a unit on NCA2 and PSS2 relative to Profile 1. Profile 3 reflects a net three quarters of a scale point improvement on NCA2 and PSS2 relative to Profile 1; Profile 4 reflects a full unit improvement on NCA2 and PSS2 relative to profile 1. The mean CR3 for Profile 2 shows a noteworthy shift from Profile 1 to a mean CR3 of 2.17, which is close to the anchor of "moderate social phobia, somewhat disabling." This indicates the amount of CR3 change that we can expect if we are able to shift *both* NCA2 and PSS2 by about half a unit in the desired direction on their metrics relative to Profile 1. For an aspirational full unit improvement for both NCA2 and PSS2 (Profile 4), we would reduce the CR3 mean even more, about half-way between the "mild social phobic" anchor and "moderate social

phobia, somewhat disabling." I sometimes find it helpful to present different multivariate profiles to help gain an appreciation of joint net effects of the mediators.

Sensitivity Analyses

Finally, there is a well-developed set of tools for OLS regression for sensitivity analyses that can be applied in OLS-based LISEM. As an example, the path coefficient linking the negative cognitive appraisals mediator to social phobia was estimated to be 0.38 (see p_4 in Table 11.4). If there are unmeasured common causes to both negative cognitive appraisals and social phobia over and above the formally modeled predictors/covariates in my model, then these confounds can artificially inflate the causal coefficient, leading the 0.38 estimate to be positively biased. In sensitivity analysis, we specify how strong the relationship between these unmeasured confounds and the two variables in question would have to be to either reduce the coefficient in question to zero if they were measured and controlled, or to reduce the coefficient in question to statistical non-significance. The program called *omitted confounders* on my website accomplishes these analyses. In the current example, I found that unmeasured confounders would have to explain more than 18.6 percent of the residual variance in negative cognitive appraisals and social phobia to reduce the causal coefficient to zero or total artifactualness.

Concluding Comments on OLS-Based LISEM

Many scientists think of structural equation modeling in terms of classic full information estimation frameworks. I have shown above that one also can implement SEM using traditional regression methods on a piecewise basis. In the social phobia example, the differences in the parameter estimates between the two approaches were minor. This will not always be the case. For example, if there are non-trivial amounts of measurement error in the measures of constructs, sizeable differences in results can occur depending on the patterning of that error, model complexity, and the data itself. One way I take advantage of LISEM is to use it as a substitute for FISEM when my sample sizes are small and cannot sustain FISEM. Also, there are analytic tools available within an LISEM framework that are not available in FISEM. I personally think it is good to have both sets of tools in your statistical toolbox for the analysis of RETs.

LISEM typically yields unbiased estimates of the parameters of interest in an RET, assuming its statistical assumptions are met or that it is robust to violations of those assumptions. What it lacks relative to FISEM is the coherence of estimates within a broader multivariate system. For example, in FISEM, the estimated total effect of the program on the outcome has an explicit mathematical relationship to the path coefficients linking the programs to the mediators and the mediators to the outcome. By contrast, such

regularities do not necessarily hold in LISEM, but the estimates can still have desirable statistical properties (e.g., consistency, unbiasedness). I personally am sometimes willing to sacrifice the mathematical elegance of FISEM if it helps me get to where I need to be when evaluating a program. See Chapter 8 for comparisons of FISEM versus LISEM.

I now apply other LISEM methods from Chapter 8 to our example. In the interest of space, I do not take you through the methods in the depth I covered FISEM or OLS-based LISEM. Rather, I focus primarily on the correspondence between the different estimates of paths p_1 through p_9 in the core influence diagram. In all cases except MIIV-SEM, I use the single indicator clinician rating, CR3, as my outcome. Also with the exception of MIIV-SEM, I do not address model fit because the method for evaluating fit by testing independence assumptions described for OLS-based LISEM applies to each approach.

LISEM: Quantile Regression

In Chapter 8, I discussed uses of quantile regression for RETs. One use is to re-focus the analysis on outlier-resistant medians rather than means. This can be useful when the outcome is subject to outliers that can distort means, such as income. Although this was not the case for the social phobia example, I nevertheless illustrate the use of quantile regression to analyze medians from an LISEM perspective. For a discussion of median regression as applied to mediation, see Yuan and MacKinnon (2014).

Quantile regression is robust to outliers for the outcome but not to unusual leverages in the predictor space, per my discussion in Chapters 6 and 8. It is good practice to test for large leverages for each estimated equation that uses quantile regression for RET analysis. I did so for Equations 11.11 through 11.14 using the program on my website called *leverage analysis*. As an example, when I applied the method to the predictors in Equation 11.14, I identified 6 cases with high leverages. I then estimated the Equation using median-focused quantile regression both with and without the high leverage cases and the results were comparable, so I concluded that the few high leverages that were present were not problematic.

Table 11.5 presents the results of the analysis of medians for the four RET equations side-by-side with the results from the prior FISEM analysis. I used conditional quantile regression to parallel the conditional-based approach of the FISEM analyses. The results were comparable, except the direct effect of the treatment on the outcome (p_7) was statistically non-significant for medians. Technically, because I am estimating different parameters (the mean versus the median), the results can differ. I also must keep in mind that (1) the FISEM analysis of means tends to have more statistical power than the analysis of medians, (2) the FISEM analysis adjusts for measurement error via the latent variables but quantile regression does not, and (3) the FISEM analysis takes into

account more information about social phobia by using three indicators.

Table 11.5: LISEM Analysis of Medians

	Original FISEM		*Analysis of Medians*	
Parameter	Coefficient	CR	Coefficient	CR
T → NCA2 (p_1)	-0.60 ±0.14	8.71*	-0.60 ±0.18	6.58*
PSS2 → NCA2 (p_8)	-0.46 ±0.10	9.52*	-0.45 ±0.20	7.75*
T → PSS2 (p_2)	1.17 ±0.10	23.64*	1.24 ±0.11	22.73*
T → ELC2 (p_3)	0.02 ±0.16	0.31	0.08 ±0.17	0.98
PSS2 → ELC2 (p_9)	-0.34 ±0.10	6.71*	-0.37 ±0.11	6.52*
T → SP3 (p_7)	-0.49 ±0.27	3.58*	-0.28 ±0.47	1.18
NCA2 → SP3 (p_4)	0.39 ±0.19	4.10*	0.36 ±0.29	2.49*
PSS2 → SP3 (p_5)	-0.71 ±0.20	7.11*	-0.93 ±0.31	5.89*
ELC2 → SP3 (p_6)	0.00 ±0.18	0.02	-0.05 ±0.29	0.37

Notes: CR=critical ratio; SP=social phobia (latent variable for FISEM, clinician rating otherwise); T= treatment condition; NCA=negative appraisals; PSS=perceived social skills; ELC=external locus of control; * $p < 0.05$

To estimate (a) the total effect of the program on the outcome, (b) the effect of the program on mediators that are impacted by other mediators, and (c) the effect of mediators on the outcome where the target mediator is impacted by other mediators, you can use the direct regression methods discussed for OLS-based LISEM. There is evidence that the combined coefficient method with Monte Carlo confidence intervals works well with quantile regression, but this needs further exploration (Shen et al., 2014). Imai et al. (2010) describe a semi-parametric approach to quantile combined coefficients.

In Chapter 8, I described the use of quantile regression to explore quantile treatment effects (QTEs). Traditional regression and FISEM test for treatment-control differences in the central tendency of an outcome or a mediator, such as its mean or median. However, it is possible for a program to have differential effects in the lower, middle, or upper portions of a distribution (see Chapter 8 for details). I used conditional quantile regression in the quantile regression program on my website to calculate the total effect of the intervention on the outcome using the direct regression method as applied to the

deciles of the social phobia outcome, CR3, predicted from the treatment condition, the baseline social phobia variable, biological sex, and hypercritical parents. Table 11.6 presents the treatment minus control quantile differences for the deciles.[6] The QTEs are comparable across each decile, suggesting the total effect of the program is roughly uniform across the social phobia distribution.

Table 11.6: QTEs for Total Effects for Social Phobia

Decile	QTE (Difference)	Critical Ratio
0.10	-1.62 ±0.21	15.31*
0.20	-1.71 ±0.23	14.38*
0.30	-1.64 ±0.29	10.98*
0.40	-1.69 ±0.32	10.55*
0.50	-1.81 ±0.35	10.29*
0.60	-1.82 ±0.33	10.78*
0.70	-1.83 ±0.37	9.82*
0.80	-1.89 ±0.28	13.14*
0.90	-1.95 ±0.33	11.53*

* $p < 0.05$

I can use quantile regression to perform significance tests of coefficient differences for any pair of quantiles. For example, the treatment minus control difference for the 0.10 decile was -1.62 and for the 0.90 decile it was 1.95. The difference is 1.95-1.62 = 0.33, which was statistically non-significant (critical ratio = 1.72, $p < 0.09$).

In addition to such analyses on the outcome per se, I also can perform them on the program effects on each mediator. A strength of LISEM quantile regression is that you can conduct a full-fledged RET analysis on any portion of an outcome distribution using both conditional and unconditional perspectives, as appropriate (see Chapters 6 and 8).

LISEM: Robust Regression

In Chapter 8, I discussed forms of robust regression that are outlier resistant, that do not assume normal distributions nor variance homogeneity. Table 11.7 shows results for the social phobia RET data using trimmed mean and MM regression. MM regression generally is resistant to both outliers and large leverages but trimmed mean regression is

[6] There are other approaches I can use to calculate quantile treatment effects and these are elaborated in Chapter 8.

only outlier resistant. I conducted the same type of leverage analysis as I did for quantile regression and did not find extreme leverages to be problematic. These models estimate different parameters (trimmed means and M estimated means rather than arithmetic means), so their results can differ. However, it is reassuring when the results converge.

Table 11.7: LISEM with Robust Regression Models

	Original FISEM		*Trimmed Means*		*MM Regression*	
Parameter	Coefficient	CR	Coefficient	CR	Coefficient	CR
T → NCA2 (p_1)	-0.60 ±0.14	8.71*	-0.62 ±0.17	7.42*	-0.61 ±0.14	8.76*
PSS2 → NCA2 (p_8)	-0.46 ±0.10	9.52*	-0.44 ±0.11	8.10*	-0.44 ±0.09	9.36*
T → PSS2 (p_2)	1.17 ±0.10	23.64*	1.21 ±0.12	20.18*	1.20 ±0.10	23.24*
T → ELC2 (p_3)	0.02 ±0.16	0.31	0.07 ±0.18	0.82	0.06 ±0.16	0.62
PSS2 → ELC2 (p_9)	-0.34 ±0.10	6.71*	-0.36 ±0.11	6.26*	-0.35 ±0.10	6.72*
T → SP3 (p_7)	-0.49 ±0.27	3.58*	-0.45 ±0.37	2.36*	-0.39 ±0.33	2.38
NCA2 → SP3 (p_4)	0.39 ±0.19	4.10*	0.35 ±0.23	3.02*	0.37 ±0.22	3.24*
PSS2 → SP3 (p_5)	-0.71 ±0.20	7.11*	-0.85 ±0.26	6.43*	-0.84 ±0.21	7.90*
ELC2 → SP3 (p_6)	0.00 ±0.18	0.02	-0.05 ±0.21	0.24	-0.03 ±0.22	0.25

Notes: CR=critical ratio; SP=social phobia (latent variable in FISEM, clinician rating for others); T= treatment group; NCA=negative cognitive appraisals; PSS=perceived social skills; ELC=external locus of control; * $p < 0.05$

To estimate (a) the total effect of the program on the outcome, (b) the effect of the program on mediators that are impacted by other mediators, or (c) the effect of mediators on the outcome where the target mediator is impacted by other mediators, you can use the direct regression analogs to those discussed for OLS-based regression. The combined coefficient method coupled with Monte Carlo confidence intervals does not apply for MM regression and trimmed mean regression.

LISEM: Bayesian Regression

Earlier I applied full information Bayesian estimation to the social phobia example. In this section, I show you how apply Bayesian LISEM to Equations 11.12 to 11.15, but on a per equation basis with single indicators; that is, I use Bayesian programming to estimate 4 separate models each consisting of one equation. This approach is not as statistically coherent as Bayesian FISEM but it can help protect against specification

error per Chapter 8. It also might be of interest in small sample scenarios where it is not possible to test complex Bayesian FISEM models (see Chapter X). Table 11.8 shows the results for the classic FISEM, the Bayesian FISEM and the LISEM Bayes methods.

Table 11.8: MLR and Bayesian Parameter Estimates

	FISEM MLR		*FISEM Bayes*		*LISEM Bayes*	
Parameter	Coef	95% CI	Coef	95% CI	Coef	95% CI
T → NCA2 (p_1)	-0.60	-0.73 to -0.46	-0.60	-0.75 to -0.45	-0.60	-0.75 to -0.45
PSS2 → NCA2 (p_8)	-0.46	-0.55 to -0.36	-0.46	-0.55 to -0.36	-0.46	-0.55 to -0.36
T → PSS2 (p_2)	1.17	1.08 to 1.27	1.17	1.07 to 1.27	1.17	1.07 to 1.27
T → ELC2 (p_3)	0.02	-0.13 to -0.18	0.02	-0.12 to 0.17	0.02	-0.12 to 0.17
PSS2 → ELC2 (p_9)	-0.34	-0.43 to 0.24	-0.34	-0.43 to -0.24	-0.34	-0.43 to -0.24
T → SP3 (p_7)	-0.49	-0.76 to -0.21	-0.49	-0.78 to -0.20	-0.42	-0.74 to -0.10
NCA2 → SP3 (p_4)	0.39	0.20 to 0.58	0.39	0.21 to 0.57	0.38	0.18 to 0.58
PSS2 → SP3 (p_5)	-0.71	-0.90 to -0.51	-0.71	-0.91 to -0.50	-0.77	-1.00 to -0.54
ELC2 → SP3 (p_6)	0.00	-0.18 to 0.18	0.00	-0.99 to 0.19	-0.03	-0.24 to 0.18

Notes: Coef=coefficient; CI=confidence/credible interval; SP=social phobia latent variable or clinician rating for LISEM; T=treatment group; NCA=negative cognitive appraisals; PSS=perceived social skills; ELC=external control

To estimate (a) the total effect of the program on the outcome, (b) the effect of the program on mediators that are impacted by other mediators, and (c) the effect of mediators on the outcome where the target mediator is impacted by other mediators, you can use the direct regression analogs to those discussed for OLS-based regression. However, in the Bayesian context, the combined coefficient method coupled with Monte Carlo confidence intervals does not apply.

LISEM: Bollen's MIIV-SEM

The MIIV-SEM approach by Bollen (see Chapter 8) is applicable to latent variable models. I applied it to the full version of the latent variable social phobia RET logic model in Figure 11.2 using CR3 as the reference indicator. MIIV-SEM is implemented in R using lavaan notation for model specification. Table 11.9 presents the syntax. I number

the lines for reference but the numbers are not included in the syntax per se. All text on the same line that follow a # are treated as comments and ignored by R. My website provides resources for programming in lavaan.

Table 11.9: R Syntax for MIIV-SEM

```
1. #input the data
2. dat<-read.table('c:/ret/socphob.txt', header=TRUE)
3. dat[dat==-999]<-NA # set missing data to NA
4. library(MIIVsem) #load library
5. #specify model using lavaan format
6. model<- '
7. # define latent variables
8. LSP1 =~ CR1+SPAI1+SPIN1
9. LSP3 =~ CR3+SPAI3+SPIN3
10. # define equations
11. NEGAPP2 ~ TREAT+PSKILLS2+NEGAPP1+HYPER+SEX
12. PSKILLS2 ~ TREAT+PSKILLS1+HYPER+SEX
13. EXTERN2 ~ TREAT+PSKILLS2+EXTERN1+HYPER+SEX
14. LSP3 ~ TREAT+NEGAPP2+PSKILLS2+EXTERN2+LSP1+HYPER+SEX
15. '
16. # end model specification
17. fit<-miive(model=model,data=dat,missing='listwise') # do the analysis
18. summary(fit) # show the output
19. fit$coefCov # show asymptotic covariance matrix
```

Lines 2 reads in the data from the file socphob.txt (see my webpage under the *program* tab and the button *how to read data into R* for how the data should be organized in the file). Line 3 sets missing data to the internal R code for missing data, NA. Line 4 opens the MIIV-SEM library. Lines 6 to 16 specify the model equations using lavaan syntax as described on my webpage, but I omit parameter labels that are typically included. Line 17 executes the analysis and stores the results in a variable called `fit`. You can use any variable name you want but is best to omit special characters from them. I rely on the default options in the program. The 'missing' option specifies how to treat missing data. The alternative to `listwise` is to specify `twostage` which invokes a method based on EM algorithms; see the MIIV-SEM manual for details. Line 18 prints out a summary of the results. Line 19 prints out the asymptotic covariance matrix.

Model Coefficients

Table 11.10 presents the estimated structural coefficients for the model for the original FISEM and for the MIIV-SEM analysis. The results are comparable. MIIV-SEM also reports factor loadings, although I do not report them here. They also were in accord with

those of the traditional FISEM.

Table 11.10: Results for MIIV-SEM

	Original SEM		*MIIV-SEM*	
Parameter	Coefficient	CR	Coefficient	CR
T → NCA2 (p_1)	-0.60 ±0.14	8.71*	-0.60 ±0.15	7.96*
PSS2 → NCA2 (p_8)	-0.46 ±0.10	9.52*	-0.46 ±0.10	9.31*
T → PSS2 (p_2)	1.17 ±0.10	23.64*	1.17 ±0.10	23.66*
T → ELC2 (p_3)	0.02 ±0.16	0.31	0.02 ±0.26	0.18
PSS2 → ELC2 (p_9)	-0.34 ±0.10	6.71*	-0.33 ±0.21	3.21*
T → SP3 (p_7)	-0.49 ±0.27	3.58*	-0.45 ±0.32	2.77*
NCA2 → SP3 (p_4)	0.39 ±0.19	4.10*	0.36 ±0.20	3.62*
PSS2 → SP3 (p_5)	-0.71 ±0.20	7.11*	-0.76 ±0.22	6.73*
ELC2 → SP3 (p_6)	0.00 ±0.18	0.02	-0.04 ±0.21	0.40

Notes: CR=critical ratio; SP=social phobia (latent variable for original SEM, clinician rating for others); T= treatment condition; NCA=negative cognitive appraisals; PSS=perceived social skills; ELC=external control, *p < 0.05

To estimate (a) the total effect of the program on the outcome, (b) the effect of the program on mediators that are impacted by other mediators, and (c) the effect of mediators on the outcome where the target mediator is impacted by other mediators, you can use the combined coefficient method coupled with Monte Carlo confidence intervals per the program provided on my webpage for Monte Carlo confidence intervals. You will need the output from Line 19 in Table 11.9 to obtain the asymptotic covariance matrix to do so (see Chapter 8 and the instructions in the program video).

Model Fit

The MIIV-SEM program provides localized tests of model fit in the form of **Sargan tests** (Kirby & Bollen, 2009). These tests are available for equations in the model that are overidentified. When an equation is overidentified, there are some parameters within it for which there are two or more ways to solve for the values of the parameters based on the observed covariances, means and variances. The different solutions should yield

comparable values if the model is correct. The Sargan test evaluates the extent to which this is the case. The test has the form of a chi square statistic that evaluates the overidentification constraints for each equation. It can only be applied to equations that are over-identified. MIIV-SEM automatically identifies instrumental variables for you, so you need not worry about doing so. The degrees of freedom for the Sargan test equals the excess number of instrumental variables above the minimum needed for identification. The null hypothesis is that all overidentification constraints of the equation hold. The alternative hypothesis is that at least one of the constraints does not hold. Rejection of the null hypothesis is evidence that the model has ill fit (see Bollen 2019, 2021 for details). A statistically significant Sargan test suggests model misspecification. For strategies to pinpoint the locus of the misspecification, see Bollen (2021).

Table 11.11 presents the results of the Sargan tests for the social phobia model in Figure 11.2. None of the tests were statistically significant, which is consistent with a good fitting model. If one or more of the tests are statistically significant, some researchers apply an FDR or Holm modified Bonferroni procedure to adjust for the multiple contrasts. Bollen & Maydeu-Oliveres (2007) propose a supplementary global chi square statistic based on the Sargan tests. One also can apply the independence test strategy described above for OLS-based LISEM to evaluate model fit.

Table 11.11: Sargon Tests for Social Phobia Example

Equation	Chi Square	df	p value
LSP1 → SPAI1	9.83	12	0.63
LSP1 → SPIN1	13.57	12	0.33
LSP3 → SPAI3	6.68	12	0.88
LSP3 → SPIN3	13.66	12	0.32
Equation 11.1	9.88	6	0.13
Equation 11.2	5.34	5	0.37
Equation 11.3	2.04	4	0.73
Equation 11.4	6.5	4	0.16

Concluding Comments for MIIV-SEM

MIIV-SEM is an interesting approach to model estimation in RETs. Its statistical engine

is a form of two-stage least squares for instrumental regression. It can readily be applied to latent variable models but retains many of the advantages of LISEM. It is robust to many forms of specification error (see Bollen, Gates & Fisher, 2018). For good introductory expositions of the approach, see Bollen (2019, 2021).

CAUSAL MEDIATION ANALYSIS

As noted in Chapter 8, Pearl's Structural Causal Modeling (SCM) approach and the causal mediation framework offer non-traditional definitions of total effects, direct effects and indirect/mediated effects with respect to mediation analysis. It turns out that when the model is based in a linear system with all continuous mediators and continuous outcomes and there are no interaction effects, the definitions SCM uses for these effects are equivalent to those yielded by traditional SEM. The social phobia example in this chapter meets these conditions. The causal mediation framework typically is applied to situations without latent variables, without certain forms of correlated disturbances, without causal relationships among mediators, and only to single mediator models. Work is ongoing to extend it beyond these contexts. I consider some of this work in future chapters (see, for example, the work on interventional indirect effects in the document on omnibus indirect effects on my webpage). However, the approach adds little to what I have covered in this chapter for continuous mediators and outcomes.

SPECIFICATION ERROR AND RESULT GENERALIZABILITY

Throughout the above modeling exercises, I have ignored a form of specification error that I routinely check in all my research, namely whether results that I observe are driven by a specific subgroup of individuals or if I am being misled by group heterogeneity in causal effects. For example, the program effect on perceived social skills was to increase the perceived social skills by approximately 1.0 units on its -3 to +3 disagree-to-agree metric. Perhaps in reality the program is quite effective for non-Latinos whereby it raises their perceived social skills by 2.0 units but ineffective for Latinos in that it does not raise their perceived social skills at all, i.e., 0 units. The average of these two effects is 1.0, which is what I observed in my analyses. If this dynamic operates, I might inadvertently be encouraging the conclusion that my program works for both groups about equally well because I ignored the ethnic difference in my modeling when, in fact, the program is more limited in scope in that it works for one subgroup (and better than I thought!) but not another.

As another example, I found that the coefficient linking negative cognitive appraisals to social phobia was about 0.40; for every one unit that negative cognitive

appraisals increases, social phobia as rated by the clinicians on their 0 to 5 rating scale tends to increase by 0.40 units. Based on these results, I concluded that this mediator was meaningful for social phobia. Suppose the causal effect for non-Latinos actually was 0.80 and for Latinos it was 0. By blithely mixing these causally heterogeneous subpopulations, I make the erroneous conclusion that the mediator is relevant to both groups when, in fact, it is not. Indeed, I mischaracterize both subgroups because I underestimate the importance of the mediator for non-Latinos and overestimate its importance for Latinos.

The above are examples of specification error caused by ignoring interaction effects or moderated relationships in the data. I conduct applied research with adolescents and there are four moderators that I routinely check to detect such specification error, (a) age of the adolescent (b) biological sex, (c) ethnicity, and (d) social class. I want to be certain that my conclusions hold across subgroups defined by each of these variables and that I am not going to mislead others by implying homogeneity of effects that are not there. Strategies to implement such specification checks are, of course, the topic of moderation analysis, which I consider in Part III of this book.

In mediation analysis, one type of specification error related to moderation that has received recent attention is that of a treatment by mediator interaction effect. Some have argued that such effects are common (VanderWeele, 2015, 2016), but there has not been empirical verification of this. I have worked in contexts where such interactions are not theoretically plausible and others where they are. Like other types of specification error related to moderation, I make it routine practice to evaluate empirically the possible presence of treatment by mediator interactions. I defer until Part III of this book consideration of how to gain perspectives on this matter.

CONCLUDING COMMENTS

Mediation analyses in RETs address three key questions, (1) does the program meaningfully affect the primary outcome(s) of interest, (2) are the mediators that a program targets indeed relevant to the outcome(s), and (3) does the program meaningfully affect the targeted mediators. Traditional mediation analysis that focuses only on omnibus indirect effects through different mediational chains does not adequately address these questions. Full information SEM can be used to explore each of the three questions and does so in a modeling framework that yields coherent coefficient estimates across multiple equations. It readily accommodates multiple mediators, causal relationships among mediators, confounder control, correlated disturbances, and latent variables that permit adjustments for measurement error, all while taking into account the extra information yielded by multiple indicators. Many popular mediation approaches

cannot accommodate the complexities that RET mediation analyses demands. SEM frameworks do so.

Although impressive on many levels, the quality of FISEM estimates is dependent on a correctly specified model and reasonable approximations to modeling assumptions. Limited information structural equation modeling can be used when the application of FISEM is questionable; for example, when sample sizes are too small for FISEM or the model is too complex or model assumptions cannot be accommodated, LISEM offers alternative analytic strategies. The present chapter applied concepts from previous chapters to the analysis of an RET on social phobia, illustrating a range of analytic tools that one can employ. Interestingly, most of the LISEM approaches I applied yielded comparable conclusions to the FISEM approach. I am not arguing that all of these methods are interchangeable. They are not. However, the results drive home the fact that there are occasions where LISEM and FISEM conclusions can converge and in cases where you may have doubts about FISEM (e.g., small sample sizes in the face of a complex model), LISEM might be an effective alternative strategy.

As discussed in Chapter 10, many researchers argue that analytic methods for RETs must be pre-specified and pre-registered before data are collected, usually committing to a single method of analysis a priori. This orientation can detract from good statistical practice in which we analyze data from multiple vantage points in the spirit of sensitivity frameworks and where analytic choices can be data driven. Rather than committing to a single tool from our statistical toolbox and being forced to use that tool no matter what, I prefer to evaluate RET data using diverse tools that reveal different nuances in the data. These multiple methods, of course, can be pre-specified.

In this chapter, I addressed the three RET questions using traditional FISEM, Bayesian FISEM, OLS-based LISEM, quantile regression, robust regression, Bayesian LISEM, and Bollen's LISEM approach. I am not saying you need to apply all of these methods to a given RET. Some methods will suit your purposes better than others and my goal is mainly to diversify the toolbox you bring to RET analysis. The key to data analysis is to be thorough, transparent, and to pursue analyses with integrity.