18

Introduction to Moderation Analysis

To generalize is to be an idiot

- WILLIAM BLAKE

INTRODUCTION

Suppose you have an illness and are considering taking a new drug for it. You learn the drug was tested in a randomized trial and that a statistically significant mean difference on the outcome was found between the group that took the medication versus a control group that received treatment as usual (TAU). Based on this average treatment effect, you know the medication was beneficial for at least some of the patients; but what you really want to know is whether the medication will be effective *for you*. You might ask yourself if the people who participated in the trial and who showed recovery/gains were like you. If the randomized trial was conducted on people who are not at all like you, this might raise doubts in your mind about the applicability of the results to you. If the people who failed to show treatment gains are like you, this also might raise doubts in your mind. As researchers identify subsets of people for whom the medication worked well, subsets for whom it worked only moderately well, and subsets for whom it did not work at all, it becomes useful to know the defining characteristics of these groups. Doing so lays the foundation for what is known as **personalized medicine**, where a medical treatment protocol is tailored to people who are most likely to benefit from that protocol; rather than treat the disease, we treat the person who has the disease by taking into account essential characteristics of the person in addition to the disease characteristics.

Moderation analysis evaluates the generalizability of program effects across different subgroups and/or settings. Is a program more effective for males than females?

Is the amount of change a program produces in a particular mediator the same for middle school students as it is for high school students? A program might seek to reduce anxiety by strengthening social support networks of patients. Does social support impact anxiety the same amount and in the same ways for everyone? If not, why would we give a program to people for whom social support has little effect? Moderation analyses in RETs address such questions. In this chapter, I build a case for conducting moderation analysis in RETs. I show how correlating variables with change scores to identify predictors of treatment success (i.e., moderators) is subject to artifacts. I consider ways we typically parameterize moderation dynamics in RETs to avoid these kind of traps when seeking correlates of changes. My focus is on continuous outcomes but I expand on moderator analysis for binary, nominal and ordinal outcomes in future chapters. I describe ways of graphing moderation and then develop the implications of what are known as ordinal and disordinal moderation. Finally, I discuss fundamental issues for asserting the absence of moderation of program effects, i.e., for asserting treatment effect uniformity. My discussion is somewhat eclectic rather than integrative because I need to lay foundations for future chapters. Be patient.

MODERATION ANALYSIS IN RETs

An advantage of moderation analysis in RETs is that it can pinpoint parts of the broader causal system that are responsible for differential program effects and potentially why those differential effects occur. By doing so, the RET provides clues for how to improve a program. Figure 18.1a illustrates moderation dynamics for a traditional RCT in which it is found that a program to reduce recidivism in inmates is less effective for inmates who have committed more serious offenses. Although this is important information, it is incomplete compared to what can be learned from an RET that combines mediation and moderation into its design. Suppose one of the mediators the program addresses is to help inmates obtain a GED (a high school degree equivalent) prior to their release from prison. The idea is that having a GED will make inmates more employable which, in turn, should reduce recidivism. Figure 18.1b illustrates four possible moderator dynamics in an RET that includes offense severity and GED in the causal analysis. First, it might be found that offense severity moderates the effect of the program on GED completion (see paths *a* and *c*). If this is the case, then I need to try to figure out why this occurs so I can address it. Why is it that inmates with severe offenses are less likely to obtain a GED after program exposure than inmates with lesser offenses? Note that the RCT without the mediation analysis would not reveal this more targeted dynamic; without the mediation analysis, I only learn that the program does not work well for people with higher offense severity.

The RET is more informative because it pinpoints that the reason this occurs is because the program is not effective at helping inmates with serious offenses obtain a GED. The question then becomes, why is this the case and what can we do about it?

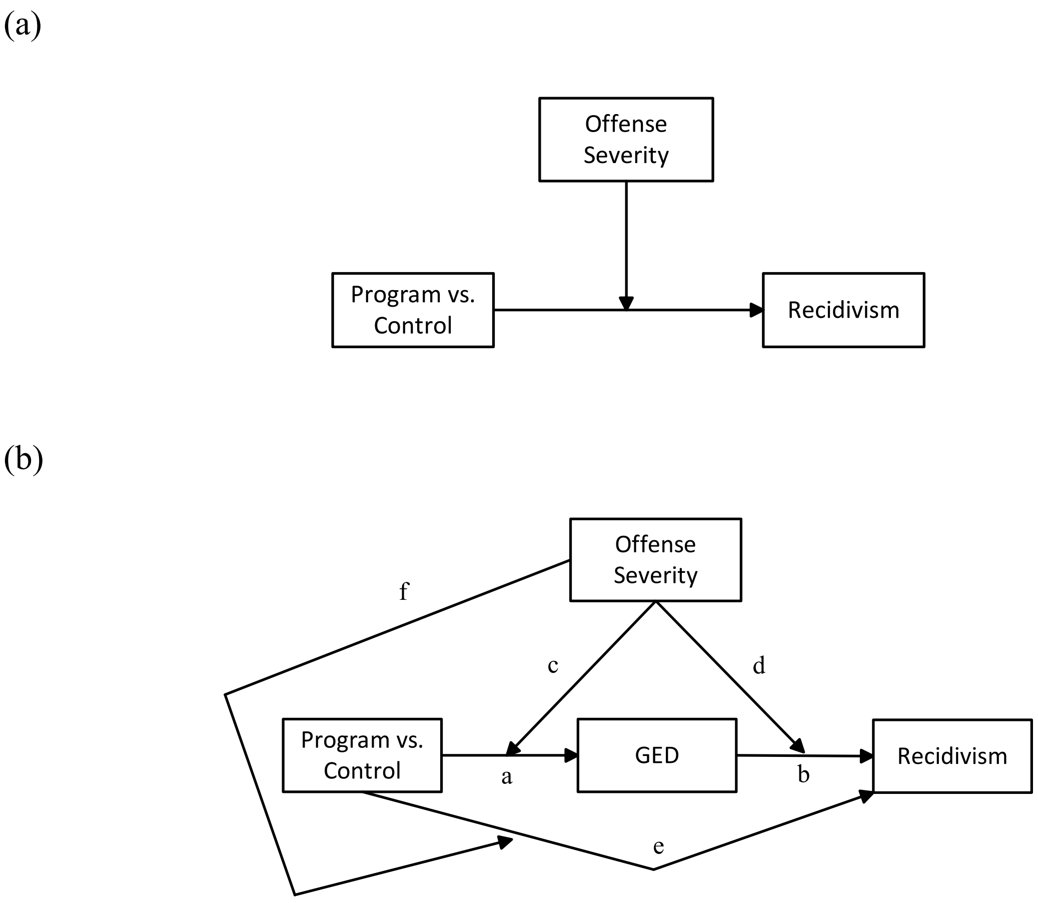


FIGURE 18.1. Moderation analysis in (a) a traditional RCT and (b) an RET

A second possible moderation dynamic is finding that the program is equally effective in helping inmates with minor and serious offenses to obtain a GED (i.e., path *c* is functionally zero), but that offense severity moderates the effect of obtaining a GED on recidivism per paths *b* and *d* in Figure 18.1b. If this result were to occur in the RET, then this tells me that the reason the program is less effective for inmates with more severe offenses lies not in the program failing at the level of helping inmates obtain a GED but rather because having a GED does not prevent recidivism for those with serious offenses on their record to the same extent that it does for inmates with more minor offenses on their record. I need to figure out why this is the case and do something about it. Again, the RCT that ignores the GED mediator does not provide me with information about where to focus my efforts to determine why the program is not working for inmates with more serious offenses on their record. The RET does.

A third possibility for moderation dynamics is that offense severity moderates *both* links in the mediational chain (paths *c* and *d*). This also is important feedback because it tells me that I need to figure out why each of the links is weaker for inmates with severe offenses. If I only fix one of the links for inmates with severe offenses, the program will still be ineffective for them because the other link remains “broken” for them. Both moderation links must be addressed simultaneously because addressing either one alone still renders the program ineffective. The traditional randomized trial that ignores mediation and moderation misses this dynamic.

A final possibility is that offense severity moderates neither of the mediational links for GED; that is both paths *c* and path *d* are functionally zero. This informs me that the source of moderation of offense severity on program effects on recidivism lies elsewhere, namely in paths *e* and *f*, not with GED. I learn not to waste time and resources trying to address the GED links as a source of differential effects for inmates with minor versus severe offenses. Rather, the moderation effect resides in a different unmeasured mediator that is subsumed within path *e*.

The RET-based results are far more informative than a simple statement from an RCT that the program is more effective at reducing recidivism for inmates with less serious offenses. Given this, I urge you to include the analysis of moderation dynamics in conjunction with mediation dynamics in your RET, as appropriate and feasible.

We can take the above RET analysis a step further by adding to the model what is known as **mediated moderation**. Consider path *c* in Figure 18.1b. Suppose we find that this link is supported such that the intervention is more effective at helping inmates with minor offenses to obtain a GED than inmates with serious offenses. I might ask myself why this might be the case. Perhaps based on past literature and/or qualitative research, I hypothesize that those who commit more severe offenses have poorer stress management skills and that these lowered skills, in turn, interfere with completion of the program activities for obtaining a GED. This dynamic is shown in Figure 18.1c where I have added stress management skills as a mediator of the moderating effect of offense severity on path *a*., i.e., I have added mediated moderation. If this reasoning proves to be correct, I might suggest to program designers to consider instituting a stress management program as a prerequisite to inmates participating in the GED acquisition component of the intervention, i.e., we can negate variation in the identified mediator of the moderation and render that moderation moot. When designing my RET, it obviously behooves me to include a measure of stress management skills so I can test this possibility.

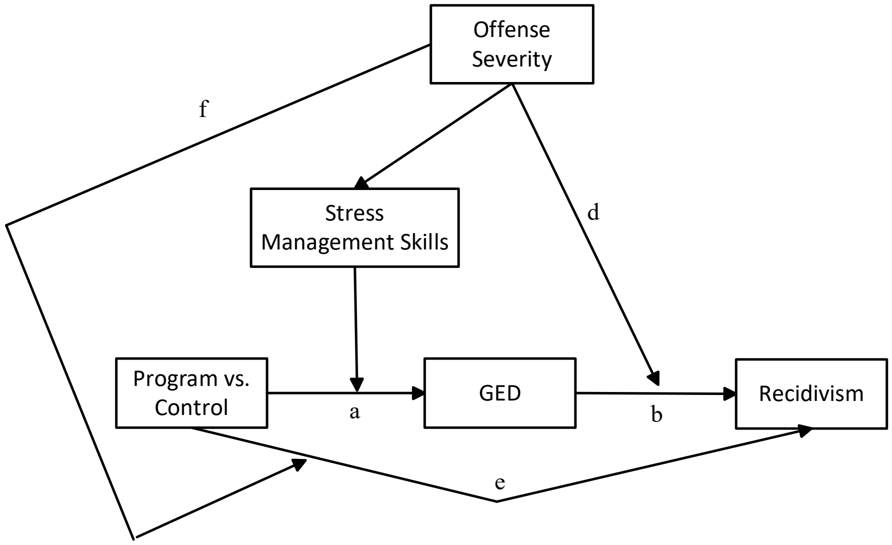


FIGURE 18.1c. Moderation analysis in (a) a traditional RCT and (b) an RET

In sum, moderation analyses are key to establishing the generalizability of program effects across different subgroups and contexts. However, they can be far more powerful and informative when they are combined with mediation analysis in the context of an RET. This further underscores my assertion in previous chapters that RETs should be our "gold standard," not more simple randomized trials.

WHEN CHANGE DOES NOT REFLECT TREATMENT RESPONSE: IMPLICATIONS FOR MODERATION ANALYSIS

Suppose we provide patients with a new medication to reduce flaring episodes for psoriasis. For Patient A, the number of such episodes across a 3 month period reduces from 12 to 6. For Patient B, the number of episodes remains constant at 10, the same rate before the patient started the medication. Many clinicians would conclude that the medication was effective for the first patient but not the second patient. As I discussed in Chapter 4, such logic is flawed. This is because the scenario is analogous to conducting an experiment using a simple pretest-posttest design for which there are confounds that can contaminate results and undermine conclusions. As examples, perhaps the medication was equally effective for the two patients but Patient A just happened to experience fewer stressors between baseline and follow-up compared to Patient B. Since stress tends to impact psoriasis flare-ups, the differential results between the two patients could be due to differences in stress that they experienced across the 3 months. Perhaps if Patient B had not received the medication, B might have seen an *increase* in flaring episodes due to an idiosyncratic increase in stress. In this case, the lack of change in flare-ups for Patient B actually reflects the medication doing its job for Patient B just as well as Patient A. Or, perhaps Patient A just happened to experience a decline in psoriasis flare-ups by chance across this particular time period, with the medication itself not really having much effect, just as it did not have much effect for Patient B.

In Chapter 4, I discussed a dozen or so artifacts that can undermine the interpretation of a change score as reflective of treatment response. To make conclusive treatment response attributions and to control for such artifacts, scientists do not rely on simple pre-post designs. Scientists know that such designs are flawed. Instead they use randomized designs by comparing people in the treatment condition with those in the control condition. Everyday medical doctors do not have this luxury and have to deal with the confounds through other means.

One practice sometimes used by researchers to identify moderators of treatment response in scientific research is to correlate individual difference variables with posttest minus pretest change scores for those who have been exposed to the intervention. For example, I might correlate biological sex with such change scores to determine if females tend to respond better to treatment than males. A problem with this strategy is that the change scores do not just reflect response to treatment; they are confounded with testing effects, history effects, instrumentation change, regression to the mean, maturation effects, experimental mortality, and selection effects, among other things (see Chapter 4 for a discussion of these phenomena). When I correlate biological sex with change, I am not only correlating it with people's treatment response but I am also correlating it with all of these other possible artifacts. This makes conclusions ambiguous.

An often unrecognized problem with this correlational approach is that the magnitude of the correlation of an external variable, Z, with a change score can be impacted by the degree of correlation between the external variable and the *baseline* outcome score independent of Z's relationship to change. As an example, research has shown that biological sex tends to be correlated with depression; females tend to report higher levels of depression than males. It turns out that given a correlation between biological sex and depression at baseline (Y_0), biological sex also will show an artifactual correlation with the change score $C = Y_1 – Y_0$. This is because Y_0 is part of C. Males will tend to show larger "improvement" than females as diagnosed by the correlation between biological sex and change in depression, but this correlation is an artifact of the correlation of sex with depression at baseline. The situation is analogous to part-whole dynamics in psychometrics. If I form a total score by summing three items, each item will show a correlation with the total score because it is part of the total score; these are what we call part-whole correlations. Analogously, Y_0 will show a correlation with the change score C because it is part of C. When we form C, we are essentially summing Y_1 with an opposite signed Y_0 to yield the value of C. Part-whole dynamics operate.

The artifactual link between Y_0 and C has been referred to as **mathematical coupling** (Tu & Gilthrope, 2007). The direction and magnitude of the correlation between Y_0 and C based on part-whole dynamics is a function of the size of the correlation between Y_0 and Y_1 and the variances of Y_0 and Y_1 (see Goldsmith et al., 2021). Here is the relevant formula that characterizes the underlying dynamics:

$$r_{Y0,C} = \frac{(r_{Y0,Y1})\sqrt{k}-1}{\sqrt{1+k-2\sqrt{k}(r_{Y0,Y1})}} \qquad [18.1]$$

where $r_{Y0,C}$ is the correlation between Y_0 and Y_1-Y_0, $r_{Y0,Y1}$ is the correlation between Y_0 and Y_1, and *k* is the variance of Y_1 divided by the variance of Y_0. If the variances of Y_0 and Y_1 are equal so that *k* = 1 and if the correlation between the baseline Y and the posttest Y (i.e., $r_{Y0,Y1}$) equals 0, then the correlation between Y_0 and the change score will be a rather substantial -0.71 just by virtue of the fact that Y_0 is part of C. If an external variable Z (like biological sex) is correlated 0.5 with Y_0 under such a scenario, then it will be correlated (-.71)(0.50) = -0.36 with change, again simply by virtue of the fact that Y_0 is part of C. In practice, the value of $r_{Y0,Y1}$ will not be zero but probably closer to 0.50, which would yield a correlation of Y_0 and C of -0.50 when *k* = 1.

My general point is that mathematical coupling also makes correlations between external variables and change scores difficult to interpret. This not only includes scenarios of correlating external variables to change, but it also includes cases where you correlate people's baseline Y scores to change in Y. Do not be surprised if your baseline measure of Y is correlated with "response to treatment" as measured by a change score because baselines scores are part of the change score. A substantial portion of this correlation likely is driven by part-whole artifacts.

A better approach to identifying moderators of treatment response is to not use change scores. Rather, we should analyze how treatment versus control differences in Y_1 vary as a function of an external variable, Z and by doing so formally bring the control group into the analysis. I now show how to parametrize moderation effects using a framework that formally compares treatment and control groups and that protects against many of the artifacts described above. I develop these ideas more in future chapters.

PARAMETERIZING MODERATED RELATIONSHIPS

To study moderation in a quantitative RET, we need to have a numerical index of it. For example, if the index equals 0, there might be no moderation at work and as the index departs from 0, more moderation is operative. In this section, I describe methods for quantifying moderation so that we can study it in an RET.

How one thinks about moderation differs depending on whether the moderator is nominal or quantitative/continuous and whether the focal independent variable is nominal or quantitative/continuous. I consider four cases here, (1) a nominal moderator and a nominal focal independent variable, (2) a nominal moderator and a continuous focal independent variable, (3) a continuous moderator and a nominal focal independent variable, and (4) a continuous moderator and a continuous focal independent variable.

Moderator Contrasts for a Nominal Moderator and a Nominal Focal Independent Variable

The key to quantifying moderation is the concept of a **single degree of freedom moderation contrast**. I illustrate the essence of such contrasts using a 2X2 factorial design with nominal variables, but the core logic applies to regression contexts as well, as I show later. Consider the following table that evaluates if a program to increase life satisfaction of older adults (scored 0 to 10 with higher scores indicating greater life satisfaction) is differentially effective for Whites versus non-Whites. The study uses a randomized trial with two conditions (program versus control). The posttest mean life satisfaction scores in the two groups as a function of ethnicity are:

	White	Non-White
Program	μ_1	μ_3
Control	μ_2	μ_4

where μ_1 is the posttest mean life satisfaction for Whites who participated in the program, μ_2 is the posttest mean life satisfaction for Whites in the control group, μ_3 is the posttest mean life satisfaction for non-Whites who participated in the program, and $_{24}$ is the posttest mean life satisfaction for non-Whites in the control group. When structuring such tables, it is common to place the levels of the presumed cause (or the independent variable) as rows and levels of the moderator as columns. There are three contrasts of interest in this table, (1) whether and by how much the program affects life satisfaction for Whites, (2) whether and by how much the program affects life satisfaction for non-Whites, and (3) whether the program effect for Whites is different from the program effect for non-Whites, that is, does the program effect generalize across ethnicity.

The first two of these three contrasts are called **simple effects**. They test the effect of the independent variable at each level of the moderator variable, which in this case is ethnicity. The two simple effects for how the program affects life satisfaction are captured by the mean contrasts:

Whites: $\mu_1 - \mu_2$
Non-Whites: $\mu_3 - \mu_4$

Analysis of such simple effects is important because the contrasts tell us if the independent variable (in this case, the program versus the control groups) has an effect on the outcome at each level of the moderator.

The question of whether program effects on life satisfaction are different for Whites versus non-Whites is evaluated by analyzing the difference between the two simple effects:

$$MC = (\mu_1 - \mu_2) - (\mu_3 - \mu_4)$$

where MC stands for a moderation contrast. If the program effects generalize across the two ethnic groups, this quantity should equal zero. When you compare the difference between mean differences in this way, you are executing a moderation contrast because you explicitly test if the effect of the independent variable on Y differs depending on the level of the moderator variable. Such contrasts are the heart of moderation analysis. They are referred to as single degree of freedom contrasts because when we execute a test of the contrast, the numerator of the F ratio for the test has one degree of freedom.

To make the above concrete, consider these sample means for life satisfaction:

	White	Non-Whites
Program	8.00	7.00
Control	5.00	5.00

The program effect for Whites is 8.00 - 5.00 = 3.00; White program participants tend to have greater life satisfaction than White controls. Although this average difference between program and control participants is non-zero, I can only conclude there is a program effect in the study *population* if the contrast yields a statistically significant p value, which then takes into account sampling error when making a conclusion. I might find in this case that the p value is $p < 0.05$.

The program effect for non-Whites is 7.00 - 5.00 = 2.00; non-White program participants also tend to have greater life satisfaction than non-White controls, which we again formally evaluate with a significance test to take sampling error into account.

The moderation contrast directly compares these two simple effects. The sample estimate of the moderation contrast is (8.00 - 5.00) – (7.00 - 5.00) = 3.00 - 2.00 = 1.00. There appears to be a differential program effect for Whites compared to non-Whites, with the program being 1.0 life satisfaction units more effective for Whites than non-Whites. Again, I need to conduct a significance test on the contrast before I can confidently conclude there is a differential effect in the study population. Note that the moderation contrast is captured in a single number that we test against zero. This will typically be the case in RETs.

Suppose instead of a 2X2 design, I work with a 2X3 design, as follows:

	White	Black	Asian
Program	μ_1	μ_3	μ_5
Control	μ_2	μ_4	μ_6

Now there are three simple effects (SE) of interest:

SE_1 for Whites: $\mu_1 - \mu_2$
SE_2 for Blacks: $\mu_3 - \mu_4$
SE_3 for Asians: $\mu_5 - \mu_6$

and there are three moderation contrasts (MC):

MC_1: Program effect for Whites versus Blacks: $(\mu_1 - \mu_2) - (\mu_3 - \mu_4)$
MC_2: Program effect for Whites versus Asians: $(\mu_1 - \mu_2) - (\mu_5 - \mu_6)$
MC_3: Program effect for Blacks versus Asians: $(\mu_3 - \mu_4) - (\mu_5 - \mu_6)$

When exploring moderation, you should routinely identify the simple effects that are of conceptual interest to you and the moderation contrasts that are of interest to you. For a nominal moderator and a nominal focal independent variable, you will then parameterize each of them using the strategies described above. I describe methods for testing the statistical significance of the simple effects and moderator contrasts in future chapters.

Moderator Contrasts for a Nominal Moderator and a Continuous Focal Independent Variable

The prior section characterized how moderated relationships are parameterized when both the moderator and the focal independent variable are nominal. In this section, I consider how to parameterize moderation dynamics when the moderator is nominal and the focal independent variable is continuous. Suppose a program seeks to increase older adults' self-esteem (measured on a 0 to 10 metric) on the assumption that doing so increases their life satisfaction, i.e., self-esteem is conceptualized as a mediator of program effects on life satisfaction. Consider the RET in Figure 18.2 that has three mediators, one of which is self-esteem. Suppose I want to determine if the effect of self-esteem on life satisfaction varies as a function of ethnicity (White, Black, Asian). Finally, suppose the RET also includes two measured covariates, which I omit from the figure to avoid clutter and I also omit the disturbance terms for the same reason.

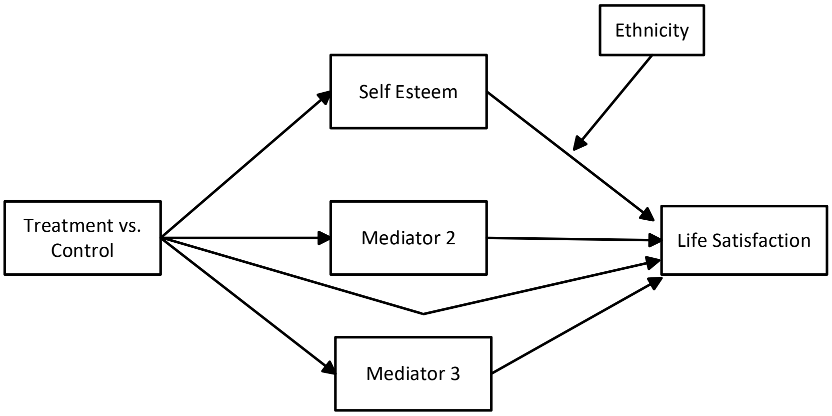


FIGURE 18.2. Example of moderated mediation

I label the path linking self-esteem to life satisfaction as p_1. Conceptually, there are three simple effects of interest, (1) the value of p_1 for Whites only; (2) the value of p_1 for Blacks only, and (3) the value of p_1 for Asians only. Consider the linear equation that regresses life satisfaction onto the three mediators, the two covariates, and the treatment condition, one of the linear equations in a standard RET analysis:

$$LS = a + p_1\,SE + p_2\,M2 + p_3\,M3 + b_1\,C1 + b_2\,C2 + p_4\,D_T \qquad [18.2]$$

where LS is life satisfaction as measured at the posttest, SE is self-esteem measured at the posttest, M2 and M3 are the other program mediators, C1 and C2 are the covariates, and D_T is a treatment dummy variable. The path coefficient p_1 is the estimated effect of self-esteem on life satisfaction holding constant the other variables in the equation. It is of interest because it estimates the causal effect of self-esteem on satisfaction independent of the other mediators and covariates, per Figure 18.2.

Suppose the sample value of p_1 for Whites is 0.60, indicating that for every one unit self-esteem increases, the mean life satisfaction is predicted to increase by 0.60 units holding the other variables in the equation constant. Suppose the p_1 value for Blacks is 0.40 and for Asians it is 0.60. The significance tests for each coefficient evaluate if each of these population simple effects is different from 0. Each test is informative. Suppose that al three are statistically significant, $p < 0.05$.

There are three moderation contrasts that follow from the simple effects. They are:

MC_1 = p_1 for Whites – p_1 for Blacks = 0.60 – 0.40 = 0.20
MC_2 = p_1 for Whites – p_1 for Asians = 0.60 – 0.60 = 0.00
MC_3 = p_1 for Blacks – p_1 for Asians = 0.40 – 0.60 = -0.20

Each of these contrasts tells us if the effect of self-esteem on life satisfaction is stronger for one ethnic group than another i.e., whether ethnicity moderates the effect of self-esteem on life satisfaction. The test of coefficient differences are single degree of freedom moderation contrasts and each is captured by a single number that we test against zero. For MC_1, we test if 0.20 is statistically significantly different from 0; for MC_2, we test if 0.00 is statistically significantly different from 0; and for MC_3, we test if -0.20 is statistically significantly different from 0. When the moderator is nominal and the focal independent variable is continuous, moderation is parameterized as differences in path coefficients for the groups defined by the moderator variable.

To summarize, when the focal independent variable is continuous (or many-valued discrete quantitative) and the moderator variable is nominal, we typically

1. Evaluate simple effect path coefficients for the focal independent variable at each level of the moderator variable (e.g., Whites, Blacks, Asians).

2. Evaluate moderation contrasts by comparing the magnitude of the simple effect path coefficients as a function of different pairs of levels of the moderator variable.

I describe methods for executing these contrasts in future chapters.

Moderator Contrasts for a Continuous Moderator and a Nominal Focal Independent Variable

In this section, I consider the case where the moderator variable is continuous and the focal independent variable is nominal or binary. Suppose the outcome is self-esteem as measured on a scale from 0 to 10 with higher scores indicating higher self-esteem. The focal independent variable is the treatment versus control condition, D_T, where 0 = the person is in the control group, 1 = the person is in the treatment group. The moderator variable is baseline stress (BS), also measured on a 0 to 10 metric in integer form with higher scores indicating higher levels of experienced stress in one's life. I want to determine if the program is more effective at changing self esteem for those with lower stress levels at baseline versus those with higher levels of stress as baseline, per Figure 18.3. The logic model is that high stress is more likely to interfere with engagement in treatment program activities which in turn, decreases the effectiveness of the program. Another way of thinking about moderation in this case is that I want to test if program effects on self-esteem generalize across the levels of stress people have before starting the program.

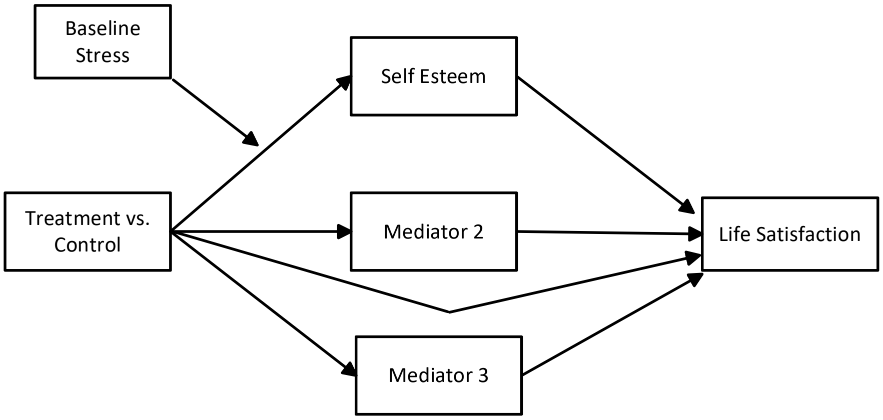


FIGURE 18.3. Example of moderation for a continuous focal independent variable

Suppose that the bulk of stress scores at baseline occur between values of 1 and 6. Table 18.1 presents a way of thinking about the simple effects and moderation contrasts. The first column lists the scores on the moderator variable, in this case, in integer form between 1 and 6. In theory, I can segregate all individuals with a baseline stress score of 1 and calculate just for them the mean posttest self-esteem for those in the control condition and the mean posttest self-esteem for those in the intervention condition. I do this in the second and third columns of Table 18.1. In the fourth column, I calculate the difference between the two means. I repeat this process in the second row but now focus only on individuals with a score of 2 on baseline stress. Table 18.1 shows the results when I continue this process on successive baseline stress scores through a score of 6.

Table 18.1. Moderation Analysis with a Continuous Moderator

Value of Moderator	Treatment Mean	Control Mean	Difference
1	7.50	3.50	4.00
2	7.00	3.50	3.50
3	6.50	3.50	3.00
4	6.00	3.50	2.50
5	5.50	3.50	2.00
6	5.00	3.50	1.50

The last column for each row in Table 18.1 represents a simple effect. Each one is the effect of the program (treatment mean minus the control mean) at a given level of the moderator variable. Usually, these simple effects and their statistical significance are of substantive interest. Because there can be so many of them with a many-valued moderator variable, we often seek ways of succinctly modeling and summarizing them. One common practice is to characterize only a few of the contrasts that span the range of the moderator values to provide readers with a sense of effect dynamics. For example, the simple effect when the moderator takes on a relatively low value (baseline stress = 1) is 4.00, when it takes on a moderate value (baseline stress = 3) it is 3.00, and when it takes on a relatively high value (baseline stress = 6), it is 1.50. These three "representative values" provide you with a sense of the simple effects as you move across baseline stress: As baseline stress gets higher, the effect of the program on the outcome weakens. I discuss in future chapters better ways of summarizing simple effects for many-valued moderator variables but the above strategy is common practice.

A moderator contrast is when we compare the effect of the program on self-esteem at one level of the moderator with the effects of the program at another level of the moderator. In the current case, there are many possible moderator contrasts because there are many levels of the moderator. I could compare the effect of the program when baseline stress equals 1 with the program effect when baseline stress equals 2 or when baseline stress equals 3, or 4, and so on. Given so many possibilities, as with simple effects, we often seek ways to succinctly model or characterize the moderation contrasts in toto so as to make the situation more manageable.

Note in Table 18.1 that there is a systematic relationship between the values of the moderator and the program simple effect reported in column 4. The relationship is linear: For every one unit the moderator variable increases, the program effect gets weaker by -0.50 self-esteem units. Based on this, I can summarize *all* of the moderation contrasts by stating this regularity in conjunction with a simple effect for the lowest value on the moderator, like this:

> *The posttest mean difference between the treatment and control conditions was 4.00 when baseline stress equaled 1.0. For every one unit that the baseline stress increased, the difference between the treatment and control group posttest means lessened by -0.50. For example, when the baseline stress was 2.0, the treatment-control difference was 3.50, which is 0.50 units weaker than when stress was 1.0; when the baseline stress was 3.0, the treatment-control difference was 3.00, a half a unit weaker than when stress was 2.0; when the baseline stress was 4.0, the treatment-control difference was 2.50, half a unit lower than when stress was 3.0. And so on.*

When a pattern of this type occurs in data (i.e., the magnitude of the effect of the focal independent variable on the outcome is a linear function of the moderator), it is given a special name; it is called **bilinear moderation**. Such moderation may or may not be common, but it turns out that, for better or worse, this form of moderation is assumed to operate by most social science researchers vis-à-vis traditional interaction modeling practices. This example represents yet another way of parameterizing moderation; one describes moderation with a single number (in this case, -0.50) that indicates how much the effect of the focal independent variable on the outcome strengthens or weakens with each unit increase in the moderator variable.

There is another way of thinking about Table 18.1 that I use in future chapters so I develop it here. If I create a dummy variable for the treatment versus control condition (D_T = 0 for control participants, 1 for program participants) and if I regress the posttest self-esteem onto D_T, the coefficient for D_T, which I signify here as *b*, is the mean difference between the treatment and control groups on self-esteem. In theory, I can calculate the value of this coefficient at each level of the moderator, which I designate as b_{m1} when the moderator equals 1, b_{m2} when the moderator = 2, and so on through b_{m6} for when the moderator = 6. For the results in Table 18.1, b_{m1} = 4.00, b_{m2} = 3.50, b_{m3} = 3.00, b_{m4} = 2.50, b_{m5} = 2.00, and b_{m6} = 1.50. I make the assumption that the values of the various b_m are some function of the moderator variable, baseline stress. For bilinear moderation, the function is assumed to be linear, which I can state in equation form as:

$$b_{mj} = a' + b' \text{ Stress}_j \qquad [18.3]$$

where *j* is a given value of the moderator, b_{mj} is the path or regression coefficient for the focal independent variable (in this case, D_T) when the moderator equals *j*, Stress_j is the value of the moderator when it has the value *j*, and a' and b' are the classic intercepts and slopes for a linear function.[1] In the current case, b' = -0.50 because for every one unit that Stress increases, the value of the coefficient for D_T decreases by -0.50 units, per the last column of Table 18.1. When analyzing bilinear moderation, a task of the analyst is to specify the values of a' and b' because they are central to moderation analysis.

Suppose instead of the results in Table 18.1, I obtained the results in Table 18.2. In this case, there is clearly moderation because the simple effects become weaker as baseline stress increases. However, the changes in b_{mj} as a function of S are not linear. For example, when Stress changes from 1 to 2, the treatment-control outcome difference weakens by -1.75, from 4.83 to 3.08; however, when Stress changes from 5 to 6, the difference weakens by only -0.10, from 1.80 to 1.70. It turns out that the function that describes how the coefficient for D_T changes as a function of Stress is quadratic, as follows:

$$b_{mj} = a' + b_1' \text{ Stress}_j + b_2' \text{ Stress}_j^2 \qquad [18.4]$$

Table 18.2. Moderator Analysis for Non-Linear Moderation

Value of Moderator	Treatment Mean	Control Mean	Difference
1	7.33	3.5	4.83
2	6.58	3.5	3.08
3	5.99	3.5	2.49
4	5.57	3.5	2.07
5	5.30	3.5	1.80
6	5.20	3.5	1.70

I delve into how one would model such **non-linear moderation** in Chapter XX. For now, the key point is that with continuous moderators we often seek to isolate a function that relates changes in the moderator to changes in the simple effects as one moves across the values of the moderator. If the function is linear, then we have bilinear moderation. If the function is non-linear, then we need to use non-linear modeling to capture the dynamics.

The concepts discussed in this section are important for moderation analysis. Faced with a continuous moderator, the typical practice is to assume bilinear moderation, i.e., to

[1] Note that this equation does not have a disturbance term. I discuss the implications of this in Chapter 19.

assume that the moderation is linear in form. It may very well not be. When you read a report, has the researcher performed diagnostics to assure you that bilinear moderation operates? Or, has the researcher simply assumed it exists? If the functional form of moderation is curvilinear, perhaps moderation is being missed in its entirety because of model misspecification. These are weighty issues and I discuss them in depth in later chapters.

Moderator Contrasts for a Continuous Moderator and a Continuous Focal Independent Variable

The final combination of variables I consider is when both the moderator variable and the focal independent variable are continuous or quantitative discrete with many values. I return to the example evaluating the effects of a program on life satisfaction (LS) through the mediator of self-esteem (SE), but now I use people's age as a moderator of the self-esteem→life satisfaction link. Age ranges from 60 to 70, inclusive, measured in integer form. I might hypothesize that the effect of self-esteem on life satisfaction weakens as age increases, per Figure 18.4.

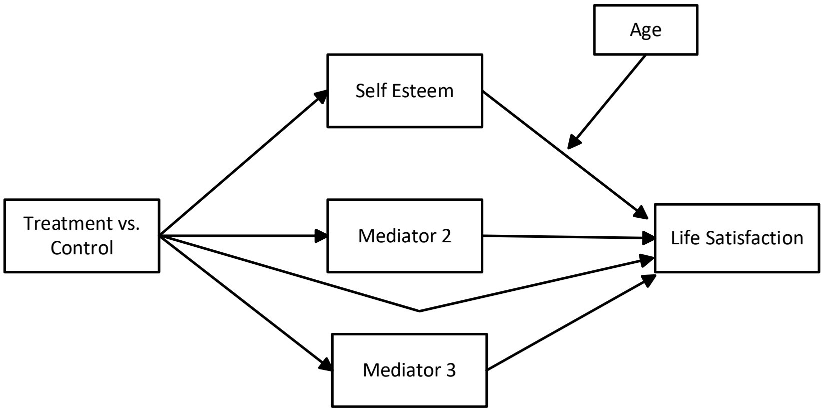


FIGURE 18.4. Example of moderation for a continuous moderator

The parameterization in this case follows closely the logic of the previous section but now instead of the regression/path coefficient for the focal independent variable being a coefficient for a dummy variable, D_T, it is a regression/path coefficient for a continuous predictor. Here, again, is the equation researchers typically would use in an RET to estimate the effect of self-esteem on life satisfaction:[2]

$$LS = a + p_1\, SE + p_2\, M1 + p_3\, M2 + b_1\, C1 + b_2\, C2 + p_4\, D_T$$

The coefficient p_1 is the estimated effect of self-esteem on life satisfaction, holding constant the other variables in the equation. In theory, I can segregate all individuals who are 60 years old and calculate the value of p_1 for them in the above equation. This has been done in the second column of Table 18.3. I then repeat this process in row 2 but now focus only on individuals who are 61 years of age. Table 18.3 shows the results when I continue this exercise through age 70, each time focusing on a single age group.

Table 18.3. Moderator Analysis for All Continuous Variables

Age	Coefficient for Self-Esteem
60	1.00
61	0.95
62	0.90
63	0.85
64	0.80
65	0.75
66	0.70
67	0.65
68	0.60
69	0.55
70	0.50

Each of the coefficients in the last column of the table is a simple effect. Each estimates the effect of self-esteem on life satisfaction at a given level of the moderator variable, age and each is of substantive interest. The moderation contrasts compare the estimated effect of self-esteem on life satisfaction at one level of the moderator with the estimated effect of self-esteem at another level of the moderator. As with the example in the previous section, there are many possible moderator contrasts because there are many levels of the moderator. Given this, I again look for a function that relates the b_{mj} to the moderator values. It turns out the results in Table 18.3 pattern themselves in accord with bilinear moderation, so I can characterize the contrasts succinctly:

[2] I omitted the covariates from the influence diagram in Figure 18.4 to avoid clutter

> *The coefficient for the estimated effect of self-esteem on life satisfaction when age equals 60 is 1.00; for every one year that age increases, the effect of self-esteem on life satisfaction weakens by -0.05.*

For example, when age = 60 years old, b_{m60} = 1.00; when age increases by one unit to 61 years old, the effect weakens by – 0.05 to b_{m61} = 0.95; when age increases by one more unit to 62 years old, the effect weakens again by – 0.05 to b_{m62} = 0.90; and so on. If the form of moderation is not bilinear, then I would use non-linear modeling to map the function. I consider the mechanics of doing so in future chapters.

Moderated Moderation

The prior sections described how to parameterize moderation when there is a single moderator. Sometimes, we encounter scenarios where we have two moderators of a relationship, not just one. Such cases can take different forms. Figure 18.5a illustrates the case where there are two moderators, biological sex and ethnicity. Each moderator separately and independently affects the program effect on the outcome, life satisfaction. For example, the effect of the program on life satisfaction might be stronger for females than males and the program effect on life satisfaction also might be stronger for Whites as compared to non-Whites. Figure 18.5b shows a different dynamic, one known as **moderated moderation**. Moderated moderation also focuses the case of two moderators but in a specialized way, namely one of the moderators moderates the moderating effect of the other moderator. Moderated moderation is more complex than traditional moderation, but it too ultimately focuses on single degree of freedom moderation contrasts. In this section, I show how to parameterize moderated moderation using nominal moderators and a nominal focal independent variable.

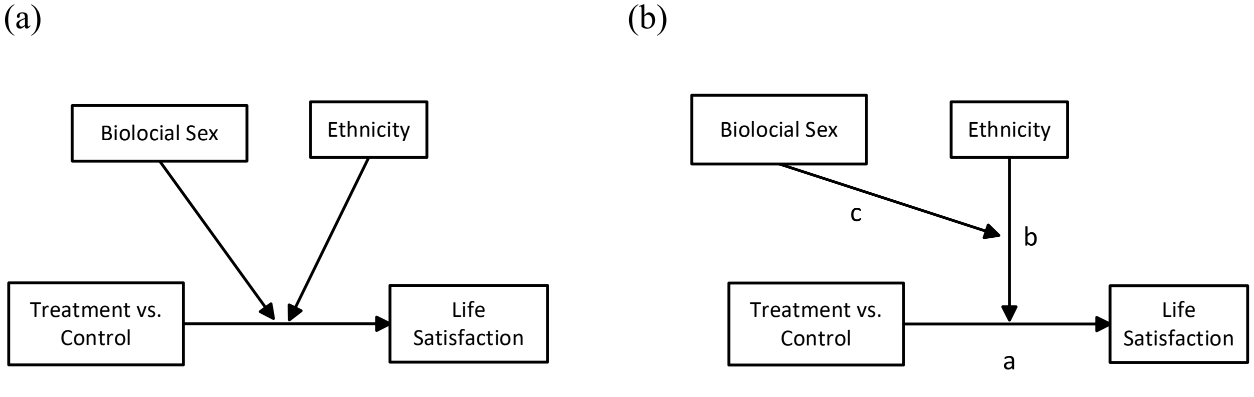


FIGURE 18.5. Example with two moderators

The example I use to illustrate parametrization is the original 2X2 design for the effects of a program on life satisfaction as a function of ethnicity but it adds biological sex as an additional moderator. For the diagram in Figure 18.5b, the treatment condition (intervention versus control) is called the **focal independent variable** and it is thought to influence the outcome. As before, ethnicity is conceptualized as a moderator variable that potentially impacts the strength of the effect of the focal independent variable on the outcome. In moderated moderation, it is called a **first order moderator variable**. Biological sex is called a **second order moderator variable** because it is thought to moderate the moderating effect of ethnicity on the effects of the program on life satisfaction. If path *c* does not exist (i.e., if it is zero), then the moderating effect of ethnicity on path *a* is the same for males and for females. If path *c* is non-zero, then the moderating effect of ethnicity on path *a* is not the same for males as females. Moderated moderation thus tests the generalizability of the moderating effect of the first order moderator across the levels of the second order moderator.

To be concrete, there are eight cells in the factorial design. Table 18.4 presents the mean life satisfaction scores for each cell. The first and second order moderators are structured as columns, with the second order moderator represented by the uppermost columns. I again adopt the convention of using the focal independent variable as rows.

Table 18.4: Moderated moderation

	Male		Female	
	White	Non-White	White	Non-White
Program	8.00	5.00	8.00	8.00
Control	5.00	5.00	5.00	5.00

I first calculate the moderation contrast just for males using the approach outlined at the outset of this chapter. I start by calculating the two simple effects that represent the program effect on life satisfaction for Whites and the program effect for Non-Whites. They are:

Male Whites:	8.00 – 5.00 = 3.00
Male Non-Whites:	5.00 – 5.00 = 0.00

The moderation contrast is

$MC_{MALES} = (8.00 – 5.00) - (5.00 – 5.00) = 3.00 - 0.00 = 3.00$

Thus, for males, the program is 3.00 life satisfaction units more effective for Whites than for Non-Whites. Next, I calculate the moderation contrast just for females. The two simple effects that represent the program effect on life satisfaction for Whites and the program effect for Non-Whites are:

Female Whites:	8.00 – 5.00 = 3.00
Female Non-Whites:	8.00 – 5.00 = 3.00

The moderation contrast is

$MC_{FEMALES} = (8.00 – 5.00) - (8.00 – 5.00) = 3.00 - 3.00 = 0$

For females, there is no moderation; the effect of the program on life satisfaction is the same for Whites as it is for Non-Whites. I can formalize the differences in these moderated effects as a function of biological sex by calculating the difference between the two moderation contrasts:

$MC_{MM} = MC_{MALES} - MC_{FEMALES} = 3.00 – 0.00 = 3.00$

where MC_{MM} represents the moderated moderation parameter estimate. Its value tells me that the moderating effect of ethnicity for males is 3.0 units stronger than the moderating effect of ethnicity for females. Stated another way, biological sex moderates the moderating effect of ethnicity on program effects on life satisfaction. Note that once again, the moderation contrast is captured in a single number that we ultimately perform a significance test on.

In sum, for moderated moderation, we typically

1. Evaluate simple effects for the focal independent variable at all possible combinations of the first and second order moderators (e.g., we test for program effects for male Whites, for male Non-Whites, for female Whites, and for female Non-Whites).

2. Evaluate moderation contrasts between the first order moderator and the focal independent variable at each level of the second order moderator (e.g., the moderation contrast for males and the moderation contrast for females).

3. Evaluate moderated moderation by examining differences in the above two-way moderation parameters from Step 2 as a function of different levels of the second order moderator.

I discuss how to execute tests of moderated moderation in Chapter XX. I also extend how to think about moderated moderation to the case of continuous moderators and continuous focal independent variables in that chapter.

Concluding Comments on Moderation Parameterization

In sum, when we analyze moderation, we are typically interested in documenting if the focal independent variable affects the outcome at each level of the moderator variable (known as a simple effect) and whether the simple effect of the focal independent variable on the outcome varies in magnitude for different levels of the moderator variable. The latter question is documented using single degree of freedom moderator contrasts, which summarizes moderation effects in the form of a single numerical index. The way we think about simple effects and moderator contrasts will differ somewhat depending on whether the focal independent variable is nominal or continuous (or a many valued discrete variable) and whether the moderator variable is nominal or continuous.

When the moderator is continuous, a crucial question is whether the form of moderation is bilinear or non-linear. Most researchers assume the presence of bilinear interactions, but this should be formally evaluated in the data.

A given link in an RET may have multiple moderators affecting it. The multiple moderators can have independent moderating effects or they might work in conjunction with one another in the form of moderated moderation. Researchers also need to be sensitive to both types of dynamics.

MODERATION VERSUS INTERACTION

The term *moderation* is often used interchangeably with the statistical term *interaction*. I engage in this practice but some social scientists object to it. The key to understanding these objections is recognizing that interaction effects can be parameterized in different ways. My emphasis has been on an approach that emphasizes the concept of effect generalizability, namely **moderation analysis**. This conceptualization address whether the effect of a focal independent variable on an outcome variable is more or less of the same magnitude at each level of the moderator variable. There is an alternative way of thinking about interactions known as **synergistic analysis**. This conceptualization is based on the idea that the "whole is not equal to the sum of the parts" and explores the unique joint effects of two variables on an outcome as compared to the additive effects of each variable considered separately. Both moderator and synergistic approaches yield the same numerical result in terms of omnibus significance tests, but they focus on very different facets of the interaction. Although my emphasis is on moderation, you should be familiar with the basics of synergistic analysis because (a) you may find it useful for some of your applications, and (b) you will encounter statements in the literature that argue that synergistic parameterizations are the *only* way in which interactions should be referenced, a position I disagree with. I describe the synergistic approach in Appendix A because it is not central to this book. I also include in Appendix B a discussion of the symmetrical nature of interactions/moderation and ambiguities in representing interactions in influence diagrams.

GRAPHING MODERATED RELATIONSHIPS

Bar Graphs and Line Plots

Some researchers depict moderation graphically. One format uses a bar graph, which is shown in Figure 18.6 (I generated the plot using the *moderator plots* program on my website). The data are for the effect of a program (versus control) on vaccination rates and the moderator is the immigration status of study participants (not born in the United States versus born in the United States). The percent of immigrants in the treatment and control groups who were vaccinated were 35% and 30%, respectively, a difference of 5%. The corresponding percents for non-immigrants were 80% and 50%, respectively, a difference of 30%. A moderation effect is present because the 5% treatment minus control difference for immigrants is smaller than the 30% difference for non-immigrants. The horizontal axis of Figure 18.6 is the moderator variable, the vertical axis is the outcome variable, and the different bars represent the levels of the focal independent variable. For the bars, the difference in height within a group (e.g., for non-immigrants) reflects the simple effect for that group. The discrepancy in the height differential between the treatment and control groups for non-immigrants compared to immigrants is the moderator effect. I added error bars for each group to reflect the lower and upper margin of error (MOE) for each percentage. Many researchers omit the MOEs to avoid clutter, instead reporting the MOEs in the main text of their research report.

Figure 18.7 presents the same data using a line plot. Each percentage is plotted and then are connected by a line if they occur in the same group for the focal independent variable (e.g., people in the control condition). The focus shifts to the distance between the lines at a given value on the X axis. Larger distances between lines reflect larger simple effects of the treatment versus control conditions for a given group. Note, for example, that the distance between the lines for the treatment and control groups for immigrants is smaller than the distance between the lines for non-immigrants. This reflects the smaller program effect on vaccination rates for non-immigrants than for immigrants. Another way of visualizing line plots is to note if the lines are parallel as you move across the X axis. Non-parallel lines imply moderation; parallel lines imply a lack of moderation. Some researchers prefer bar graphs as visual aids for simple effect and moderation analysis and others prefer line plots. Which do you prefer?

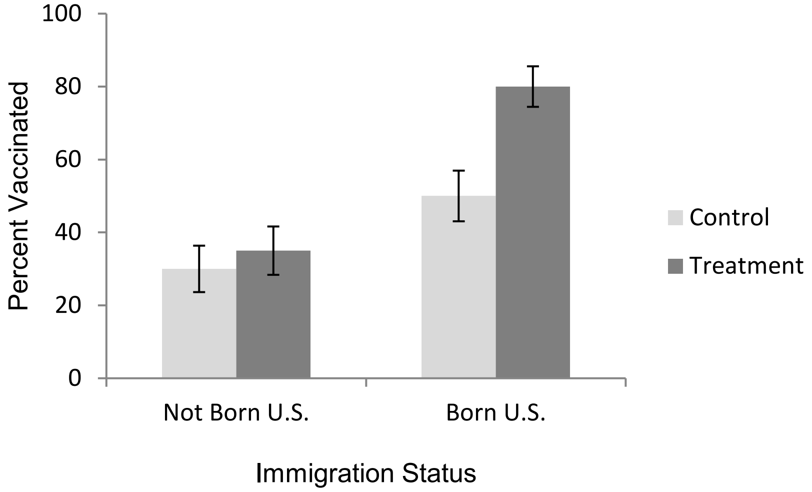


FIGURE 18.6. Bar graph of moderation

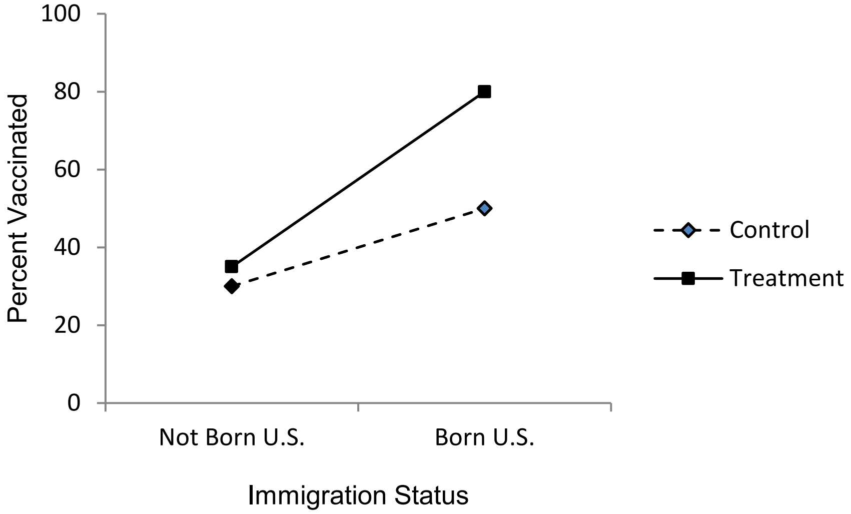


FIGURE 18.7. Line plot of moderation

Plots with Continuous Moderators

When the moderator, focal independent variable, and outcome all are continuous, it is challenging to illustrate moderation effects graphically. One strategy is to use a **surface plot**, also known as a **perspective plot**.[3] A surface plot is a three dimensional plot that shows the value of a summary statistic of the outcome (usually the mean of Y) as a function of the two continuous predictors, X and Z. Suppose that Y is a person's intention to obtain a vaccination for a disease, X is the perceived susceptibility to the disease the vaccine prevents, and Z is the perceived severity of the disease if the disease were to be contracted. Each construct is measured on a continuous -3 to +3 metric where higher scores indicate more positive intent, greater perceived susceptibility and greater perceived severity, respectively. A 0 on each metric is "neutral." Suppose that vaccination intent (labeled vintent in my plots) is a perfect additive linear function of perceived susceptibility (psuscept) and perceived susceptibility (psuscept). Figure 18.8 presents a surface plot for such data, generated using the *surface plots* program on my website. The program is interactive and allows you to reposition the surface for viewing the plot from different angles by dragging it with your mouse (watch the video associated with the program).

The essence of an additive linear function is that the surface forms a plane, which is a flat surface with no thickness. Figure 18.8 shows the plane from three different angles. In Panel 18.8a, the outcome is positioned on the left, the focal independent variable is positioned on the top, and the moderator variable is positioned on the bottom. This represents the positioning I recommend because, as you will see later, it is congenial to characterizing moderation. Panel 18.b reorients the plot somewhat to help you see the plane dynamics at work. Panel 18.8c shows an extreme angle where I position the plane to look at one edge (perceived severity) while visually manipulating the other edge (perceived susceptibility) so that the latter vanishes from view (much like holding your hand perfectly parallel to the ground at eye level but then, in this case, elevating your fingers to form a 45 degree angle). In the current case, this particular angle highlights the slope between vaccination intentions and perceived severity; higher perceived severity leads to higher intentions to obtain a vaccination.

[3] Some methodologists use the terms interchangeably and others make distinctions between them.

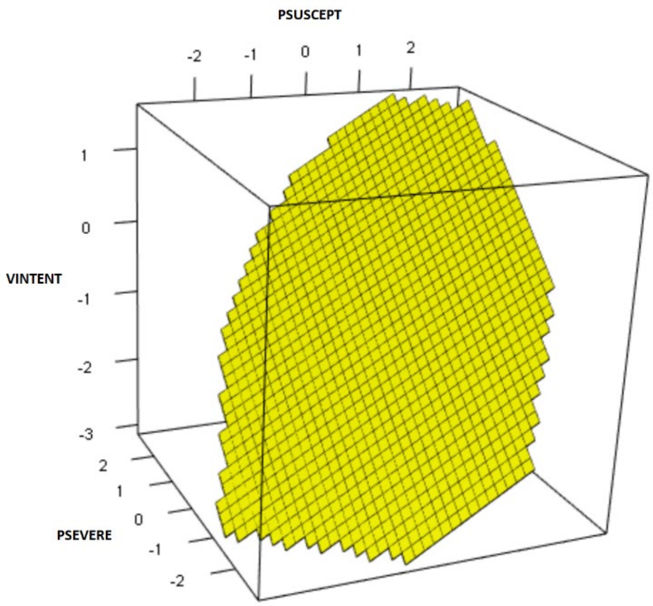


FIGURE 18.8a. Surface plot from angle 1

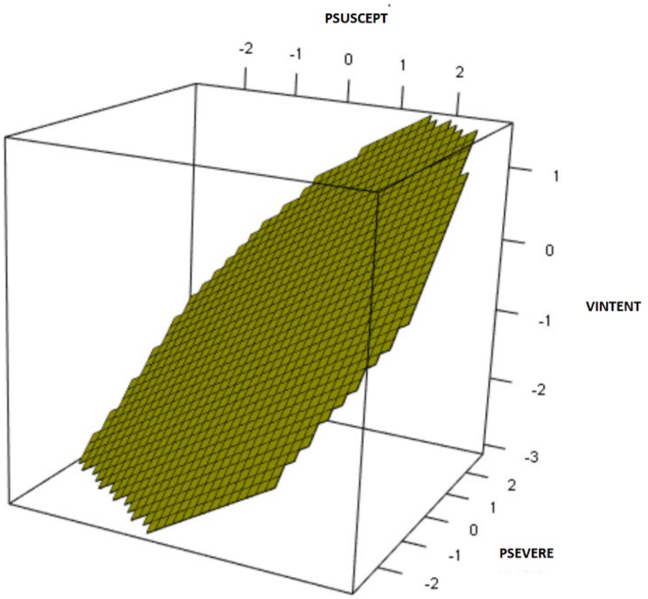


FIGURE 18.8b. Surface plot from angle 2

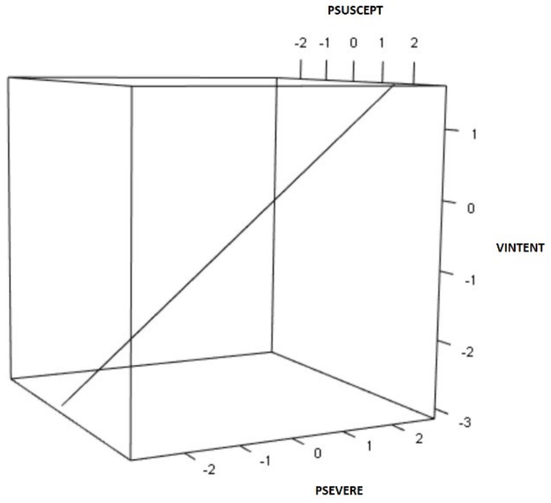


FIGURE 18.8c. Surface plot from angle 3

A surface plot for a model with moderation between predictors looks different than that for a model that does not have moderation. Figure 18.9a presents a surface plot for data that were generated from a main effect only model for the vaccination example and Figure 18.9b presents a surface plot for data that were generated from a model in which perceived severity moderates the impact of perceived susceptibility on vaccination intent, i.e., the effect of perceived susceptibility on vaccination intent becomes stronger as perceived severity increases. In Figure 18.9a, I highlight 3 "slices" of the surface by placing red lines through boxes of the grid I added to the plot at points A, B and C. Point A reflects the case where perceived severity is low, Point B where it is moderate, and Point C where it is high. Each line reflects the slope for the effect of perceived susceptibility on vaccination intent. Noteworthy in Figure 18.9a is that the slope is the same for Points A, B and C; they are functionally parallel indicating no moderation. Look at the same points in Figure 18.9b. You can see that the three slopes are non-parallel, reflecting the operative moderation dynamics.

Another strategy for plotting all continuous variables is to present the regression lines relating the continuous focal independent variable to the outcome at three or four strategically selected values of the moderator variable. Table 18.3 presented regression coefficients reflecting the effect of self-esteem on life satisfaction for 11 different age groups. I might present the regression lines in a graph for individuals who are 60, 65, and 70 years old, per Figure 18.10 using the *multiple curves* program on my website.

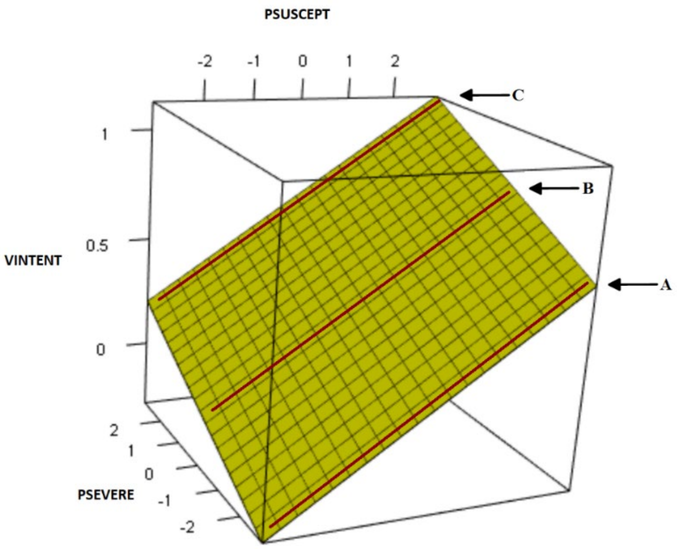


FIGURE 18.9a. Main effect model

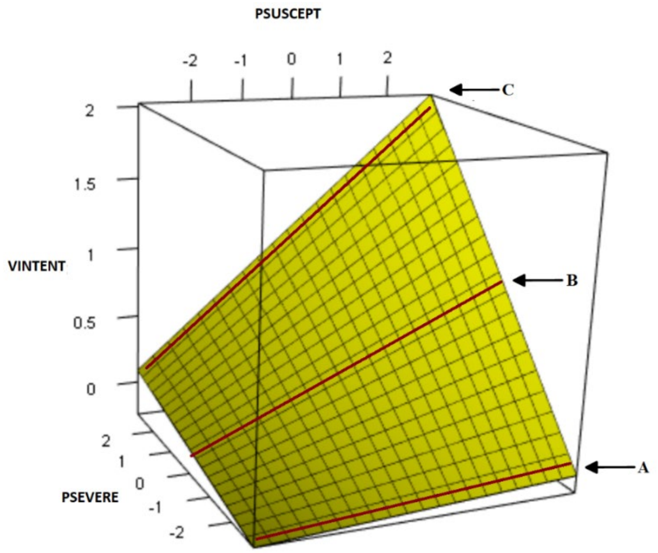


FIGURE 18.9b. Model with moderated effect

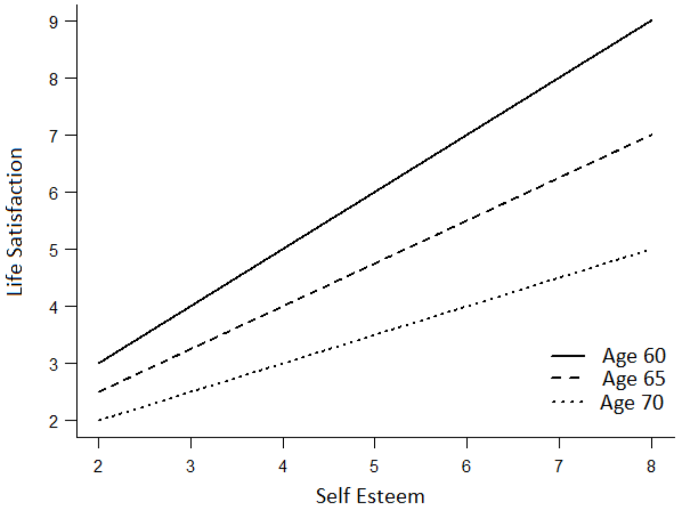


FIGURE 18.10. Line plot for different age groups

In sum, many researchers represent moderation using plots, either bar graphs, line plots, or surface plots. They can be effective ways of highlighting moderation, although I personally orient better to numbers rather than graphs.

Moderated Moderation Plots

For moderated moderation, when the first order and second order moderators as well as the focal independent variables all are categorical, researchers sometimes use **side-by-side plots** to illustrate moderation. Table 18.4 presented data that showed the moderating effects of ethnicity on the effects of a treatment program on life satisfaction. The moderating effect for males was different from the moderating effect for females. Figure 18.11 translates the data in the table to a side-by-side bar graph for purposes of illustrating moderated moderation. The different moderation dynamics as a function of biological sex are evident in the two plots. I generated this plot using the *moderator plots* program twice to generate each plot separately and then I pasted them together into the Paint program in Microsoft.

If the moderator variable is continuous, then researchers might present a side-by-side plot for two or three selected values of the continuous moderator that are substantively of interest, per Figure 18.10, but one for a low value of the moderator, one for a medium value of the moderator, and a third for a high value of the moderator, side-by-side.

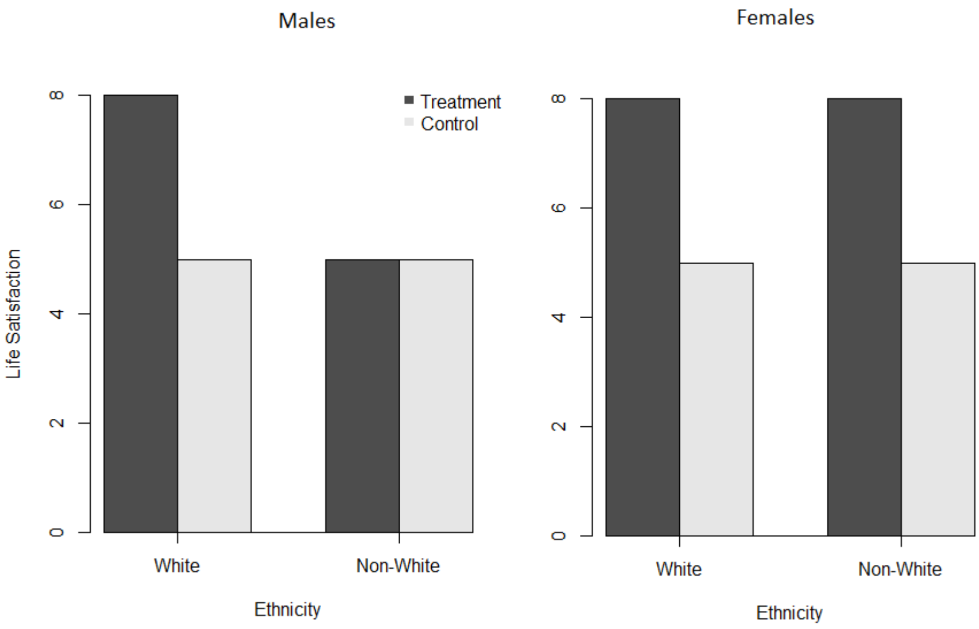


FIGURE 18.11. Side by side plot

ORDINAL AND DISORDINAL MODERATION

Consider a study where I compare two different forms of therapy, cognitive behavioral therapy (CBT) and psycho-education (PE) to foster pain tolerance by people living with chronic pain. Pain tolerance is scored on a 0 to 6 metric with higher scores indicating greater tolerance. I hypothesize that baseline anxiety moderates the relative effectiveness of the two treatment types. For individuals with low levels of baseline anxiety, CBT should be more effective because it relies on solid principles with an extensive scientific base; for individuals with high levels of baseline anxiety, PE should be more effective. This is because CBT is a more challenging form of therapy for patients to master and the higher levels of anxiety will likely get in the way of mastering CBT related tasks. Such is not the case (as much) for PE. Baseline anxiety is measured on a 15 to 45 metric with higher scores indicating higher levels of anxiety.

Social scientists distinguish between ordinal moderation and disordinal (also known as crossover) moderation. **Disordinal moderation** occurs in a line plot when the line for one group intersects the line for the other group. **Ordinal moderation** is when the lines are nonparallel, but the lines do not intersect. Figure 18.12 presents an example of each type of moderation for the pain tolerance example. I treat baseline anxiety as the moderator variable and treatment type (CBT versus PE) as the focal independent variable. I regress posttreatment pain tolerance (adjusted for pretreatment pain tolerance) onto baseline and anxiety and plot the regression lines for each treatment group on the same plot. Like any line plot, the distance between the two regression lines at any given point on the X axis reflects the relative effects of the two treatments at that point.

In Figure 18.12a, there is ordinal moderation because the regression lines do not cross. Note that no matter where I look on the X axis (baseline anxiety), CBT is superior, on average, to PE. This is because CBT is higher up on the Y axis than PE. To be sure, the superiority of CBT to PE for increasing pain tolerance is greater when baseline anxiety is low. But even when baseline anxiety is at its highest, CBT fares better than PE. The recommendation for which treatment a patient should undergo would always favor CBT over PE

Contrast this with Figure 18.12b, where there is disordinal moderation. In this case, the regression lines intersect. The point of intersection is theoretically important because it is the point on baseline anxiety dividing where CBT is more effective than PE and where PE is more effective than CBT. From the results in Figure 18.2b, if a patient has a baseline anxiety score less than 29.3, s/he likely should be treated with CBT; if the patient has a baseline anxiety score greater than 29.3, then s/he should instead be treated with PE. Isolating the point of intersection in disordinal moderation is substantively important.

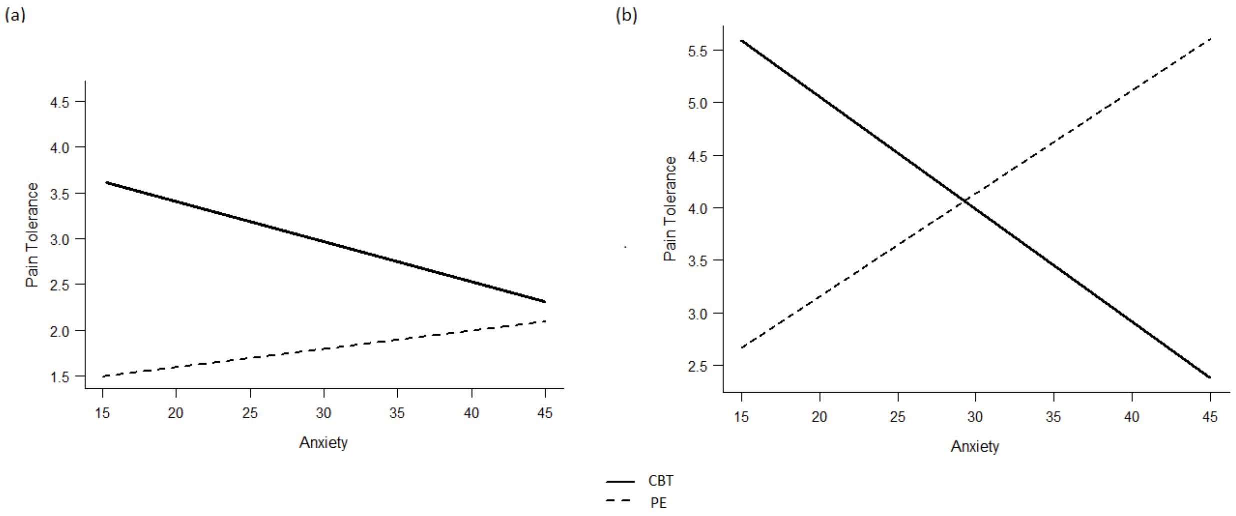


FIGURE 18.12. Ordinal and disordinal moderation

More than 75 years ago, Cronbach and Gleser (1957) reviewed the logic of classification decisions in clinical, educational, and organizational settings. They noted that decisions about the assignment of people to treatments (e.g., clinical interventions, type of educational curricula, type of job) are frequently guided by the identification of crossover points in disordinal moderation: Persons to the right of the crossover point are assigned to one treatment while persons to the left of the crossover point are assigned to the other treatment condition. By contrast, ordinal moderation implies the same treatment should be used for all individuals.

Interestingly, for any pair of nonparallel regression lines, there will always be a point where the lines intersect; if you extend the regression lines far enough, they eventually will cross over. In this sense, all interactions for moderation of this type are crossover interactions. Interactions are said to be ordinal if, *within the range of scores being studied*, the regression lines do not intersect.

How do we isolate the point of intersection for disordinal moderation? Consider the case of two groups. Let a_1 be the intercept for the first group, a_2 the intercept for the second group, b_1 the regression/path coefficient for the first group, and b_2 the regression/path coefficient for the second group. The formula for identifying the point of intersection is:

$$PI = (a_1 - a_2) / (b_2 - b_1) \qquad [18.6]$$

For Figure 18.12b, the regression equation for the PE group is 1.193 + (.098)*Anxiety the for the CBT group it is 7.913 + (-.107)*Anxiety. The point of intersection is

$$PI = (1.193 - 7.193) / (-0.107 - 0.098) = 29.27$$

This type of information is key to personalized medicine. I discuss strategies for analyzing disordinal moderation and points of intersection in Chapter XX.

ASSERTING GROUP EQUIVALENCE

It is not uncommon for researchers to assert generalizability given a statistically non-significant moderation contrast between groups. Suppose I test if the effect of a program to increase the mean monthly retirement savings is different for males as opposed to comparably paid females. The null hypothesis for the moderation contrast is

$$H_0: (\mu_{TREAT,FEMALES} - \mu_{CONTROL,FEMALES}) - (\mu_{TREAT,MALES} - \mu_{CONTROL,MALES}) = 0$$

and the alternative hypothesis is

H_1: $(\mu_{TREAT,FEMALES} - \mu_{CONTROL,FEMALES}) - (\mu_{TREAT,MALES} - \mu_{CONTROL,MALES}) \neq 0$

If the sample data yields a statistically non-significant result, I cannot conclude that the program is equally effective for males and females because this is tantamount to accepting the null hypothesis, which is inappropriate. I *can* conclude that the differential effect observed in the sample data is not sufficiently strong for me to say that moderation operates relative to biological sex; but that is different from concluding that the program effect is the same for males and females.

Per my discussion of effect sizes in Chapter 10, to make strong statements of program effect equivalence, we need to invoke equivalence standards and formally test program effects in light of those standards. In Chapter 10, I developed the notions of the latitude of meaningfulness, the latitude of no effect, and the latitude of effect ambiguity for evaluating effect sizes. In future chapters, I use a similar framework for purposes of making assertions of group equivalence vis-à-vis moderation. Specifically, I draw on a large statistical literature called **equivalence testing** that specifies equivalence thresholds that are used to assert functional equivalence of effects. For example, suppose I evaluate a program to increase monthly retirement savings and seek to test if the program effects for comparably paid males and females are similar. I might state an equivalence standard of $25 per month; if the *population* program effect for females is within ±$25 of the *population* program effect for males, I will conclude that the program effect is *functionally equivalent* for males and females. If I find evidence that this is the case, I can make a strong statement of effect generalizability for the two populations. Just as we had to tackle the difficult issue of defining effect size standards in Chapter 10, so must we tackle such issues for asserting effect generalizability.

CONCLUDING COMMENTS

I have described ways of thinking about and forming quantitative representations of moderated relationships. These representations are key to analyzing moderation in RETs. When we explore moderated relationships, we want to gain insights into the effects of the focal independent variable on the outcome at each level of the moderator variable. These are the simple effects in the analysis. We also want to determine if program or mediator effects at one level of the moderator are different from effects at another level of the moderator. I have not described statistical methods for conducting such contrasts. I do so in future chapters. For now, you should ensure you understand the logic of moderation, the different forms it can take, and the ways we parameterize moderation effects.

APPENDIX A: MODERATION VERSUS INTERACTION REVISITED

In this Appendix, I describe the difference between the moderation approach to interaction analysis and the synergistic approach. I explain the latter by first developing the statistical concepts of **treatment effects** and **residualized means** using the classic analysis of variance (ANOVA) model.[4] Consider an example with two dichotomous mediators as influencers of a continuous outcome. The example uses data that are hypothetical with equally sized subgroups to make it easier for me to convey the relevant logic. The population is a group of pregnant women who were both heavy drinkers and heavy smokers prior to becoming pregnant. There are two groups in the RET, (1) those who participated at the start of their pregnancy in a program to convince them not to smoke cigarettes or drink alcohol during their pregnancy, and (2) those in a control condition that received a non-smoking-non-drinking message about nutrition during pregnancy, i.e., it was an active control group. The outcome is the birthweight of the newborn measured in grams (1 pound is about 454 grams). Smoking was measured dichotomously (0 = did not abstain from smoking during pregnancy after program enrollment, 1 = abstained from smoking during pregnancy after program enrollment) as was alcohol use (0 = did not abstain from drinking alcohol during pregnancy after program enrollment, 1 = abstained from drinking alcohol during pregnancy after program enrollment). Table A.1 presents the average birthweight of the mother's newborn as a function of the four possible combinations of the two mediators collapsing across program condition. (G is the grand mean of newborn birthweight across all mothers).

Table A.1: Mean Birthweight as a Function of Smoking and Drinking

	Abstained from drinking	Did not abstain from drinking	Marginal Mean
Abstained from smoking	3,500	3,400	3,450
Did not abstain from smoking	3,200	3,200	3,200
Marginal Mean	3,350	3,300	G=3,325

For the moderator approach, there is an interaction effect in the data. If I use abstinence of drinking as the moderator variable, the effect of abstaining from smoking

[4] I use the term *treatment effect* here ***not*** to refer to an intervention designed to change an outcome in an RET but as a statistical term. I considered using different jargon but want to remain true to the literature surrounding synergistic conceptualizations of interactions. In this section, I use the term *program* to refer to an intervention.

on birthweight when mothers abstained from drinking is

SE_1: 3,500-3,200 = 300

The effect of smoking on birthweight when mothers did not abstain from drinking is

SE_2: 3,400-3,200 = 200

and the moderation contrast is

$MC = SE_1 – SE_2 = 300 – 200 = 100$

For the synergistic approach, there also is an interaction effect but it is framed differently using what are called main effect treatment effects and residualized means. For the main effect of smoking, consider the marginal means for abstaining from smoking. As seen in Table A.1, the mean newborn birthweight for mothers who abstained from smoking was 3,450 grams whereas the mean birthweight for everyone in the study was G = 3,325 grams. The effect of abstaining from smoking seems to be to raise the birthweight of newborns, on average, by the difference between these two means or

$\tau_{AS} = \mu_{AS} – \mu_G = 3{,}450 - 3{,}325 = 125$

where τ_{AS} is the treatment effect of abstaining from smoking, μ_{AS} is the marginal mean for abstaining from smoking, and μ_G is the grand mean.

Next, I calculate the treatment effect for the other level of abstaining from smoking vis-à-vis the same process, namely I subtract the grand mean from the marginal mean for not abstaining from smoking:

$\tau_{NAS} = \mu_{NAS} – \mu_G = 3{,}200 - 3{,}325 = -125$

where τ_{NAS} is the treatment effect for not abstaining from smoking and μ_{NAS} is the marginal mean for not abstaining from smoking. The effect of not abstaining from smoking is to lower the birthweight of newborns, on average, by -125 grams.

Note that the treatment effects for the two levels of abstinence from smoking are equal in value but opposite in sign. It turns out this will always hold for the case of a factor with two levels with equal N. It also is the case that the sum of the treatment effects across the levels of a factor equals 0. This latter property also will be true for factors with more than two levels.

A similar analysis can be conducted for the treatment effect for abstinence from drinking. Here are the treatment effects for the two levels of this variable:

$\tau_{AD} = \mu_{AD} - \mu_{G} = 3{,}350 - 3{,}325 = 25$

$\tau_{NAD} = \mu_{NAD} - \mu_{G} = 3{,}300 - 3{,}325 = -25$

The effect of abstaining from drinking (τ_{AD}) is to raise the birthweight of newborns, on average, by 25 grams and the effect of not abstaining from drinking (τ_{NAD}) is to lower the birthweight of newborns, on average, by -25 grams.

The interaction effect also is represented by treatment effects. However, to isolate them, I first need to remove the influence of the main effects from the data so that the interaction effects are not contaminated by them. This process involves focusing on a given cell mean (e.g., the cell for abstaining from both smoking and drinking, $\mu_{AS,AD}$) and literally subtracting the treatment main effect for "abstained from smoking" and for "abstained from drinking" from that mean:

$\mu'_{AS,AD} = \mu_{AS,AD} - \tau_{AS} - \tau_{AD} = 3{,}500 – 125 – 25 = 3{,}350$

where $\mu'_{AS,AD}$ is the adjusted cell mean with the main effect influences removed from it. Put into words, the effect of abstaining from smoking was to raise newborn weight by 125 grams for mothers in this particular cell; I "nullify" that effect by subtracting 125 grams from the cell mean. Similarly, the effect of abstaining from drinking was to raise newborn weight by 25 grams; I "nullify" that effect by subtracting 25 grams from the cell mean. I repeat this process for each cell of the design, which yields a set of **residualized cell means** per Table A.2. For example, for $\mu_{AS,NAD}$, I nullify the effect of abstaining from smoking (which was to raise newborn weight by 125 grams) by subtracting 125 grams from the mean and I nullify the effect of not abstaining from drinking (which was to lower newborn weight by 25 grams) by adding 25 grams to the cell mean. Note that the marginal adjusted means for the main effects now equal the grand mean in Table A.2. This is because I removed or nullified the influences of the main effects from the data.

Table A.2: Adjusted Cell Means for Birthweight

	Abstained from drinking	Did not abstain from drinking	Marginal Mean
Abstained from smoking	3,350	3,300	3,325
Did not abstain from smoking	3,300	3,350	3,325
Marginal Mean	3,325	3,325	G=3,325

An interaction treatment effect can now be defined for each cell by subtracting the grand mean from the adjusted cell mean in Table A.2, analogous to the process I used to define the treatment effects for each level of the main effects. This yields the table of interaction treatment effects shown in Table A.3. The values reflect the synergistic effects of the two factor levels that define a cell over and above the influences of the main effects. For example, the combination of abstaining from both smoking and drinking synergistically increases newborn birthweight by 25 grams over and above the additive main effects of these variables. It is these synergistic effects that Rosnow and Rosenthal say define an interaction and it is these effects, they argue, that one must reference and interpret to fully understand interaction effects.

Table A.3: Interaction Treatment Effects

	Abstained from drinking	Did not abstain from drinking	Marginal Mean
Abstained from smoking	+25	-25	0
Did not abstain from smoking	-25	+25	0
Marginal Mean	0	0	

All of the above logic is captured in the classic ANOVA model that expresses a person's score on the outcome as an additive function of the treatment effects:

$$Y_i = \mu_G + \tau_{Aj} + \tau_{Bk} + \tau_{ABjk} + \varepsilon_i \qquad [A.1]$$

where μ_G is the grand mean, τ_{Aj} is the treatment effect for group *j* that the individual is in relative to factor *A*, τ_{Bk} is the treatment effect for the group *k* that the individual is in relative to factor *B*, τ_{ABjk} is the (residualized) interaction treatment effect for factors *A* and *B* and ε is a disturbance term of the difference between the observed Y and the predicted Y.

Rosnow and Rosenthal (1989a, 1989b) believe that the correct way to characterize interactions is to focus substantive interpretation of interactions on the interaction treatment effects in Equation A.1. They are critical of scientists who use the difference between mean differences approach, i.e., the moderator framework. Their insistence has been questioned by multiple methodologists, so the topic is controversial (*cf* Becker & Coolidge, 1991; Meyer, 1991; Petty, Fabrigar, Wegener & Priester 1996; Rosnow & Rosenthal, 1989a, 1989b, 1991, 1995; Ross, & Creyer, 1993). One problem with Rosnow and Rosenthal's argument is that the ANOVA model only becomes statistically identified when certain constraints are imposed on it, such as the constraint that the treatment effects must sum to zero (see Meyer, 1991). These constraints shape the meaning of the parameters and it is not uncommon for the interaction parameter values to lack meaningful substantive interpretation in light of the mathematical constraints imposed on them. For example, in Table A.3, the synergistic effect of abstaining from both smoking and drinking is to increase newborn weight by 25 grams, which makes intuitive sense. However, the synergistic effect of smoking cigarettes and drinking alcohol during pregnancy also is to increase newborn weight by 25 grams, which is counter-intuitive. Perhaps one can devise theory to accommodate such synergy (e.g., smoking and drinking reduces stress which then positively affects birthweight), but I personally would think long and hard about the assumptions being made in the synergistic approach that yield parameter values requiring such logic.

In program evaluations, we typically want to know how group means (or percentages) differ from one another and if mean differences vary as a function of other variables. Such questions are forthrightly addressed by the moderator framework. I am just as justified to insist that interaction terminology be reserved for differences between mean differences as Rosnow and Rosenthal are to insist that it be reserved for synergistic effects. In the larger statistical literature, both synergistic and moderator characterizations of interaction analysis have rich traditions; the term "interaction" is firmly entrenched in both. Sometimes one parameterization works better than the other for the questions a researcher seeks to address. To me, the situation is analogous to choosing coding schemes for dummy variables in multiple regression to represent nominal variables. All would agree that using dummy variables to represent nominal predictors is a reasonable strategy. However, depending on one's research question, one might use effect coding, dummy coding, or orthogonal coding for the dummy variables in order that the associated coefficients (parameterizations) for them address the questions the researcher seeks to answer. Such also is the case for interaction analysis, where the interaction can be parametrized using the moderation framework, the synergistic framework, or even both frameworks sequentially to answer one's questions. For elaboration of this perspective, see Petty et al. (1996) and Meyer (1991). Parenthetically, frameworks that focus on mean contrasts rather than treatment effects in ANOVA contexts are often referred to as a **cell means approach** (Kirk, 2012, Maxwell et al., 2017). The moderator framework uses a cell means approach. The synergistic framework does not.

APPENDIX B: THE SYMMETRY OF MODERATION

In this appendix, I develop the implications of a symmetry property that exists for moderation analysis. To make my discussion concrete, I use the example from Appendix A on abstaining from smoking and abstaining from drinking during pregnancy and their estimated effects on the weight of newborns. If you have not read Appendix A, familiarize yourself with the example before proceeding.

Classic moderation involve three variables, a focal independent variable, a moderator variable, and an outcome variable. Given two predictors in a linear equation, one predictor must be assigned the role of being the focal independent variable and the other predictor the role of being the moderator. Statistically, the choice of which predictor is to take on which role does not matter, as I illustrate shortly. However, substantively, the choice does matter. Your choice should be dictated by how you want to frame the moderation in the larger substantive narrative you are weaving. Sometimes role assignment is obvious, such as when I seek to test the generality of a program effect on an outcome across ethnicity; the treatment condition is obviously the focal independent variable and ethnicity is the moderator. However, in some contexts the choice can go either way in terms of substantive justification.

Such is the case for the smoking and drinking study in Appendix A; I can either frame my narrative around the effects of abstaining from smoking on newborn birthweight as moderated by abstaining from drinking; or, I can frame my narrative around the effects of abstaining from drinking on birthweight as moderated by abstaining from smoking. Statistically, the value of the moderation contrast parameter and its significance test is the same no matter which perspective I choose. Let me show you why this is the case. I reproduce for convenience the table of means in Table B.1 for the newborn weight study.

Table B.1: Mean Birthweight as a Function of Smoking and Drinking

	Abstained from drinking	Did not abstain from drinking	Marginal Mean
Abstained from smoking	3,500	3,400	3,450
Did not abstain from smoking	3,200	3,200	3,200
Marginal Mean	3,350	3,300	G=3,325

If abstaining from smoking is the focal independent variable and abstaining from drinking is the moderator variable, the moderation contrast is

MC = (3,500-3,200) – (3,400-3,200) = 300 – 200 = 100

If abstaining from drinking is the focal independent variable and abstaining from smoking is the moderator, the moderation contrast is

MC = (3,500-3,400) – (3,200-3,200) = 100 – 0 = 100

Note that the result is identical to that of the prior framing. Stated another way for those of you familiar with the use of product terms for interaction analysis (Chapter 19), the path coefficient for the product term of abstaining from drinking times abstaining from smoking when newborn weight is regressed onto the product term and its components parts will equal 100 no matter which variable is designated as the moderator and which as the focal independent variable. As well, the t test and p value for the coefficient will be identical. It is in this sense that the choice of which variable is the moderator is arbitrary, at least from a purely statistical point of view. Note also that such symmetry is true when the predictors are continuous or when they are combinations of nominal and continuous variables. I show in future chapters how to model asymmetric moderation if that is your preference.

Suppose the report I am writing is focused on the effects of smoking on birthweight. It would then make sense that I would use abstaining from smoking as the focal independent variable when reporting the moderation effect. If, on the other hand, my report is focused on the effects of abstaining from drinking on birthweight, then it makes sense that I would use it as the focal independent variable. Sometimes I find myself in scenarios where both narratives are of interest. When this occurs, I might present the data both ways but acknowledge the contrast redundancy that is operating between them.

There is a second ramification of moderation symmetry that you should keep in mind. It concerns the representation of moderation in influence diagrams. For the newborn weight example, I can draw the influence diagram in either of two ways. The first approach (Figure B.1a) treats abstaining from drinking as the moderator variable and the second approach (Figure B.1b) treats abstaining from smoking as the moderator variable. Some students ask if a diagram like Figure B.1a can be represented as in Figure B.2a to acknowledge the additional presence of the "main effect" of abstaining from drinking on newborn weight in addition to it taking on the role of a moderator. The answer is that doing so can be misleading because it fails to acknowledge that the effect of drinking on newborn weight is different depending on whether a woman also abstains from smoking. Figure B.2a does not reflect this.[5]

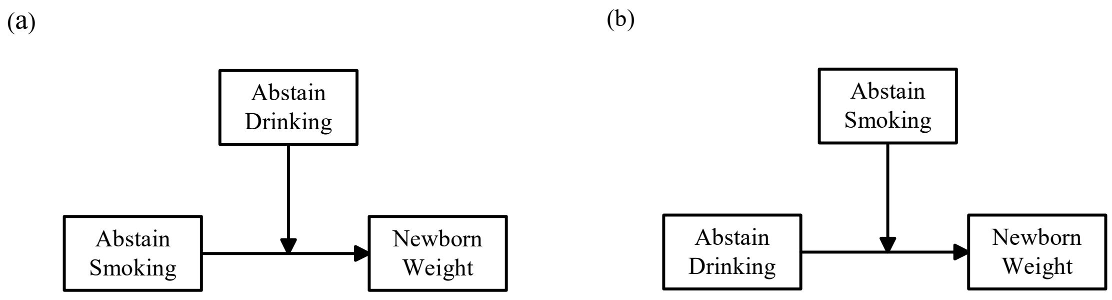


FIGURE B.1. Two ways of representing moderation

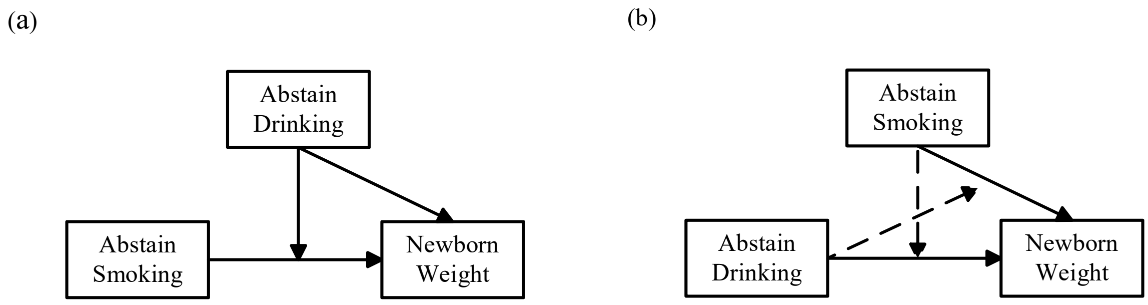


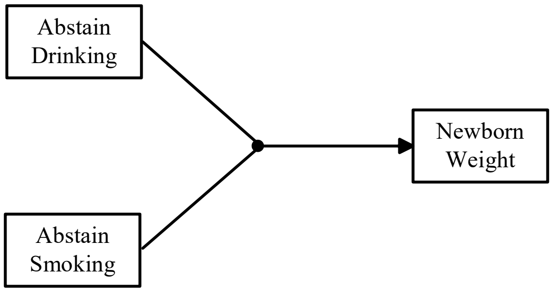
FIGURE B.2. Additional ways of representing moderation

The symmetry can be depicted using dashed arrows per Figure B.2b. Dashed arrows are used because they reflect a different dynamic than a traditional solid arrow. You then define the meaning of the dashed arrow in the text accompanying the figure or in a footnote. However, when I have used Figure B.2b in reports coupled with explanatory footnotes, reviewers often complain that it is confusing. Given this, I just draw moderation in the traditional way per Figure B.1a or Figure B.1b and trust that the reader knows about effect symmetry and understands it is implied in the diagram. However, I fully recognize the ambiguity of the practice.

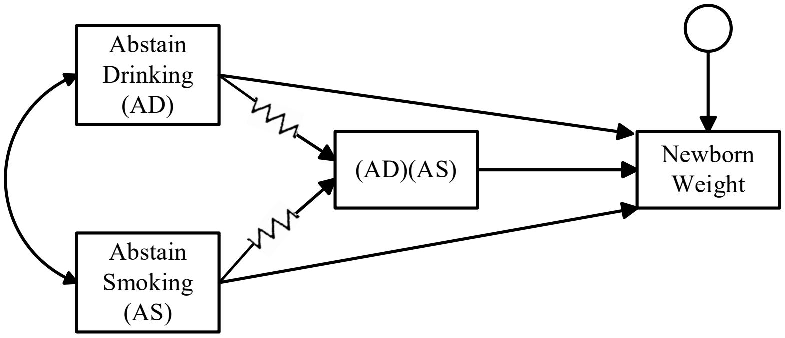
Some methodologists, including the authors of Mplus, use a schema in which lines from the interacting variables converge on a point, from which an arrow emanates

[5] I omot cirved arrows for exogenous variables and disturbance terms in all figures to avoid clutter

towards the outcome variable, like this:



Diagramming interactions in DAGs is controversial; see Nilsson, Bonander, Strömberg & Björk (2021) for recommendations, none of which have been fully accepted by the DAG community. Finally, Bollen (1995) suggests a strategy that represents the interaction in the form of a product term (see Chapter 19) as its own box that is connected by "sawtooth" arrows for its component parts:



The sawtooth arrows stand for nonlinear relations between the variables at the base of the arrow and the head of the arrow. In prior diagrams, I omitted correlations between exogenous variables and disturbance terms to reduce clutter. Here, I include them to emphasize that Bollen explicitly omits a disturbance term for the product term because it is completely determined by the two component parts.

Ultimately, an influence diagram may not be able to capture rigorously the intended way of treating interactions at both a conceptual and statistical level. One typically presents an influence diagram to provide a conceptual sense of one's logic and then adds text to explain how the interaction is parameterized and why. That is why I use the more traditional way of diagramming the interaction rather than formally incorporating the product term or a separate box for it. But, to each his/her own.