$b = r_{13} - ac$,
$c = r_{23} - ab$, and
$a = r_{12}$.

We will solve the equation for b by substituting the third and second equations (for a and c, respectively) into the first equation:

$$
\begin{aligned}
b &= r_{13} - r_{12}(r_{23} - r_{12}b) \\
&= r_{13} - (r_{12}r_{23} - r_{12}^2 b) \\
&= r_{13} - r_{12}r_{23} + r_{12}^2 b \\
b - r_{12}^2 b &= r_{13} - r_{13}r_{23} \\
b(1 - r_{12}^2) &= r_{13} - r_{12}r_{23} \\
b &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2}
\end{aligned}
$$

See if you can use the same approach to solve for c.

2 The other method of developing equations to solve for paths is called the first law of path analysis (Kenny, 1979, p. 28). The correlation between Y (a presumed effect) and X (r_{xy}) is equal to the sum of the product of each path (p) from all causes of Y times the correlation of those variables with X: $r_{yx} = \sum p_{yz} r_{xz}$. Using the first law, the correlation between Motivation and Achievement is $r_{32} = br_{12} + cr_{22}$, which reduces to $r_{32} = br_{12} + c$ (description and equation adapted from Kenny, 1979, p. 28). The advantage of the first law is that it can be used to generate equations for any type of model, whereas the tracing rule works only with simple recursive models.

3 These rules are that standardized coefficients above .05 could be considered small; those above .10, moderate; and those above .25, large. These rules apply primarily to manipulable influences on school learning.

4 Kline (2011) adds a fourth condition, that the direction of the presumed causation is correctly specified (p. 98). This is a little more specific than we want to get right now, and we will deal with this problem in the next chapter.

5 Total effects are sometimes referred to as total causal effects. It is also possible to subtract the total causal effects from the original correlation to determine the noncausal (or spurious) portion of the correlation.

12
Path Analysis
Dangers and Assumptions

Path analysis is not magic; it does not prove causality. It does not make a silk purse out of a sow's ear; it cannot turn poor data into valid causal conclusions. Like multiple regression, there are assumptions underlying path analysis and the use of multiple regression to estimate paths. Like multiple regression, path analysis is open to abuse. This chapter will discuss these assumptions and the dangers of path analysis; it will also discuss how to avoid the dangers of the method.

ASSUMPTIONS

Because we have so far been using multiple regression to estimate path models, it should not be surprising that the basic assumptions of multiple regression also apply to path analysis. As discussed in Chapter 9, these include the following:

1. The dependent variable is a linear function of the independent variables. In addition, the causal direction in the model must be correct.
2. Each person (or other observation) should be drawn independently from the population.
3. The errors are normally distributed and relatively constant for all values of the independent variable.

Multiple regression analysis assumes that the errors are uncorrelated with the independent variables or, in the jargon of path analysis, the disturbances are uncorrelated with the exogenous variables. Therefore, the causal mechanism underlying our path analysis (or multiple regression) model needs to conform to these same constraints in order for the regression coefficients to provide accurate estimates of the effects of one variable on another. This assumption also implies several additional assumptions; to the extent that the following conditions are violated, the paths (regression coefficients) may be inaccurate and misleading estimates of the effects.

1. There is no reverse causation; that is, the model is recursive.
2. The exogenous variables are perfectly measured, that is, they are completely reliable and valid measures.
3. "A state of equilibrium has been reached" (Kenny, 1979, p. 51). This assumption means that the causal process has had a chance to work.
4. No common cause of the presumed cause and the presumed effect has been neglected; the model includes all such common causes (Kenny, 1979).

If these sound a lot like the assumptions from Chapter 9, you are perceptive; they are virtually the same but rewritten in path analytic lingo. These assumptions are also required any time we wish to interpret regression coefficients in a causal, or explanatory, fashion.

The first assumption (of the second set) is really twofold. It first means that we have paths drawn in the correct direction. We have already discussed how this is done and will continue to discuss this critical issue in this and later chapters. This assumption also means, as indicated, that the model is recursive, with no feedback loops or variables both causing and affecting other variables. There are methods for estimating such models, but ordinary multiple regression is not a valid method for nonrecursive models.

The second assumption is one we can only approximate. We all know there is no such thing as perfect measurement, especially in the social sciences. When we begin discussing latent variable SEM, we will see how serious our violation of this assumption is and what can be done about it. For now, I will simply note that if our exogenous variables are reasonably reliable and valid little harm is done, meaning our estimates of effects are not overly biased.

The third assumption is that the causal process has had a chance to work. If motivation affects achievement, this process presumably takes a certain amount of time, and this time must have elapsed. This assumption applies to all causal research. Consider an experiment in which children are given some treatment and subsequently measured on a dependent variable. If you make these measurements too soon, not allowing the treatment to work, you will miss spotting any real effects your treatment may have. The amount of time needed depends on the process being studied.

The final assumption is the most crucial, and it is one we have returned to over and over in this book. We will now explore it in more depth, because the danger of omitted common causes is the biggest threat to the causal conclusions we reach from path analysis, in particular, and nonexperimental research in general. Again, I remind you that these assumptions apply to *any* explanatory use of MR.

THE DANGER OF COMMON CAUSES

Suppose I were to go into my local elementary schools and ask every student to read the Gettysburg Address, and I scored each student on the number of words he or she read correctly within 2 minutes. Suppose that I also measured each child's shoe size. If we correlate these two variables (reading skill and shoe size), we likely will find a substantial correlation

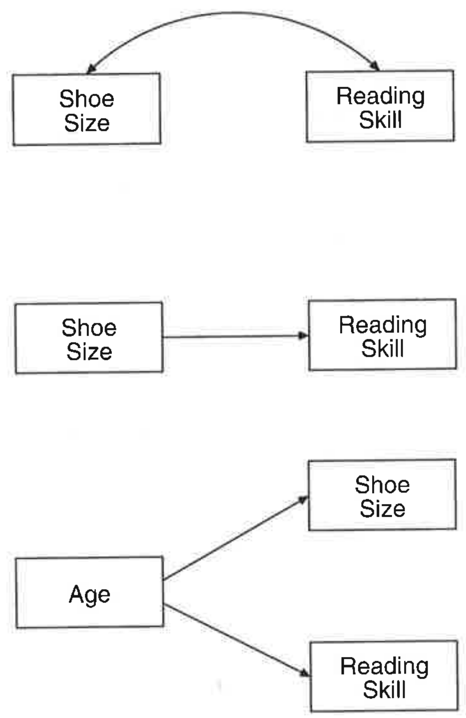

Figure 12.1 Spurious correlation in path form. Although shoe size and reading skill are correlated, shoe size does not cause reading skill, nor does reading skill cause shoe size. There is a third variable, age or growth, that affects both reading skill and shoe size. This common cause of shoe size and reading skill is why the two variables are correlated.

between them. This correlation is illustrated in the top of Figure 12.1 by the curved line between the two variables. It is foolish, however, to conclude that shoe size affects reading skill (as is done in the middle portion of the figure), and it is equally foolish to conclude that reading skill affects shoe size. The reason is that there is a third variable—age or growth—that affects both shoe size and reading skill, as symbolized by the bottom portion of Figure 12.1. Older students, on average, are larger (and thus have larger shoes) and read better than do younger students. The bottom of the figure illustrates the true causal relation among these variables; shoe size and reading skill are correlated only because the two are affected by age. The correlation between shoe size and reading skill is the essence of what we call a spurious correlation. The term *spurious correlation* means that two variables are not related by one variable affecting the other but are the result of a third variable affecting both (cf. Simon, 1954).

This example also illustrates the essence of the problem we have been referring to as that of a neglected common cause. If we set up a path analysis of the reading–shoe size data in which we assumed shoe size affected reading skill (as in the middle of Figure 12.1), the results would not tell us we were foolish, but instead would suggest that shoe size had a substantial impact on reading skill. The reason, again, is that we neglected to control for age, the common cause in our analysis. If we controlled for age, we would see the apparent effect of shoe size on reading skill diminish to zero. The model is crucial; for the estimates to be accurate, we must control for important common causes of our presumed cause and presumed effect. This problem is referred to as omitted common causes, spurious correlation, or the third-variable problem.

A Research Example

A more realistic example will further illustrate the problem. There is ample evidence that involvement by parents in education improves students' learning (Christenson, Rounds, & Gorney, 1992), but estimates of the effects of parent involvement on learning vary widely across studies. Figure 12.2 shows a plausible model of the effects of Parent Involvement on 10th-grade GPA. For this model, Parent Involvement was defined as a combination of parents' educational aspirations for their children and communication between parents and their children about school. Background variables—potential common causes of Parent Involvement and 10th-Grade GPA—include students' Ethnic origin, their Family Background characteristics, and their previous school performance (Previous Achievement). Let's concentrate on this final variable. Previous Achievement should certainly affect students' current academic performance, since it forms a basis for all future learning. But should students' previous academic performance also affect the degree to which parents are involved in students' schooling? I think it should; it should affect both parent involvement, in general, and more specifically parents' aspirations for their children's future educational attainment (one of the components of parent involvement). We could turn to previous research and determine that students' previous performance or aptitude indeed affects their parents' level of involvement. In other words, Previous Achievement, or aptitude, appears to be a likely common cause of both Parent Involvement and current GPA.

I estimated the model using the NELS data; the results are shown in Figure 12.3. Parent Involvement appears to have a moderate effect on student GPA ($\beta = .160$). The results show that our supposition about Previous Achievement was also correct: Given the adequacy of the model, the results suggest that Previous Achievement had a large effect on both GPA (.417) and Parent Involvement (.345). Previous Achievement thus appears to be an important common cause of Parent Involvement and current Grades.

What would happen if we were not attuned to the importance of students' previous school performance? What if we had not built Previous Achievement into our model? What if we had neglected this important common cause? The results of such neglect are shown in Figure 12.4. In this model, Previous Achievement was not included; this important common cause was not controlled. The result is that the model substantially overestimates the effect of Parent Involvement on GPA: the effect in this model is .293, as opposed to .160 in the

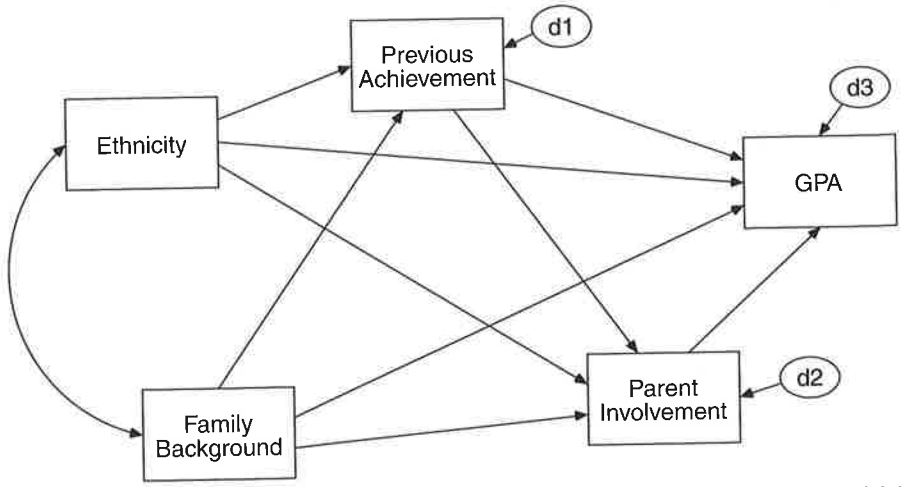

Figure 12.2 Model of the effects of Parent Involvement on high school GPA. The model is just-identified and recursive.

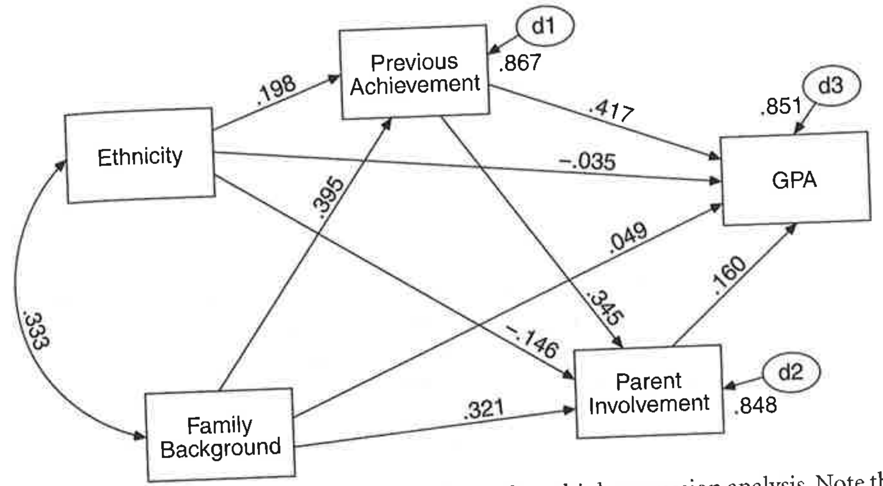

Figure 12.3 Parent Involvement model estimated through multiple regression analysis. Note the effect of Previous Achievement on Parent Involvement and GPA.

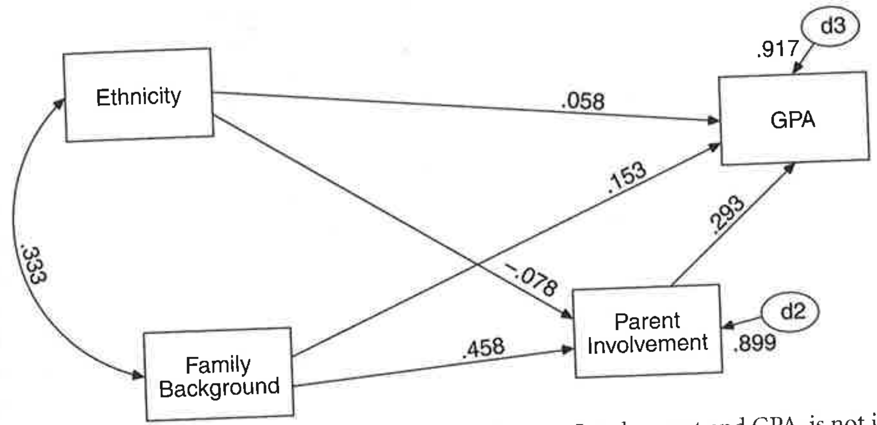

Figure 12.4 Previous Achievement, a common cause of Parent Involvement and GPA, is not included in this model. Notice the inflation of the path from Involvement to GPA.

previous model. With the omission of this important common cause, we overestimated the effect of Parent Involvement on GPA.

This example illustrates the importance of including known common causes in path models. The example also illustrates the most frequent consequence of neglecting these common causes: When a common cause is omitted from a model, we often end up overestimating the magnitude of the effect of one variable on another.[1] Finally, the example illustrates one possible reason for the variability in findings concerning the effect of parent involvement on school performance: not all research has controlled for previous achievement (and there are other possible explanations, as well). Research on parent involvement is not the only area in which researchers have likely overestimated effects by ignoring important common causes. For example, Page and Keith (1981) showed how Coleman and colleagues (Coleman, Hoffer, & Kilgore, 1981) had overestimated the effects of private schooling on student achievement by ignoring student ability as a potential common cause of achievement and private school attendance. In fact, if you are suspicious of the findings of nonexperimental research, you should probably first look for neglected common causes as the reason for misleading findings.

Note that there was nothing in the analysis summarized in Figure 12.4 that told us we had missed an important common cause. The analysis did not explode; no alarm bells went off. How then do you know that you have included all relevant common causes in your research? A good understanding of relevant theory and previous research are the keys to avoiding this deadly sin, just as they are for drawing the model in the first place.

Common Causes, Not All Causes

Unfortunately, many neophytes to path analysis (and nonexperimental research in general), terrified of neglecting a common cause of a presumed cause and a presumed effect, include every variable they can think of that might be such a common cause. Others misunderstand the admonition about *common* causes and try to include all possible causes of *either* the presumed cause or the presumed effect. Both approaches lead to overloaded and less powerful analyses (by reducing degrees of freedom in the regression), ones that are more likely to confuse than inform (and see Darlington, 1990, chap. 8, for additional dangers with including too many variables).

I demonstrated in note 2 in Chapter 4 that the inclusion of a noncommon cause in a regression does not change the estimates of regression coefficients. Here we will demonstrate this truism again using the current example. Focus again on Figure 12.3. For this model, we do not need to include all causes of Parent Involvement in the model, nor do we need to include all causes of GPA in the model. This is fortunate, because there must be hundreds of variables that affect GPA alone! All we need to include in the model are *common* causes of Parent Involvement and GPA. Note the effect of Ethnicity on Parent Involvement and GPA. Ethnicity affects Parent Involvement; other things being equal, minority students report greater involvement than do majority students (majority students are coded 1 and minority students coded 0). But once the other variables in the model are controlled, Ethnicity had no meaningful effect on GPA ($\beta = -.035$). Despite its inclusion in the model, it appears that Ethnicity is not a common cause of Parent Involvement and GPA. If my argument is correct, if variables need not be included in a model unless they are common causes, then the exclusion of Ethnicity from the model should have little effect on our estimate of the magnitude of influence of Parent Involvement on GPA. As shown in Figure 12.5, the exclusion of Ethnicity

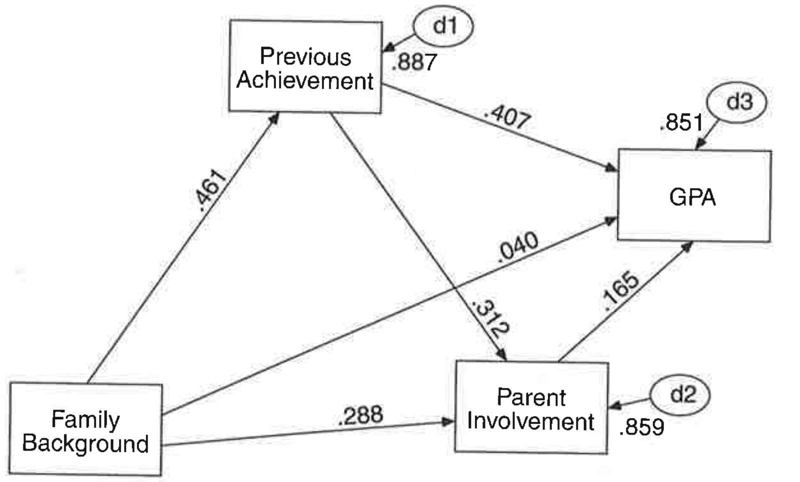

Figure 12.5 In this model, Ethnicity was excluded. But Ethnicity was not a meaningful common cause; it affected only Parent Involvement, not GPA, in Figure 12.3. Thus, its exclusion in this model has little effect on the estimate of the effect of Parent Involvement on GPA.

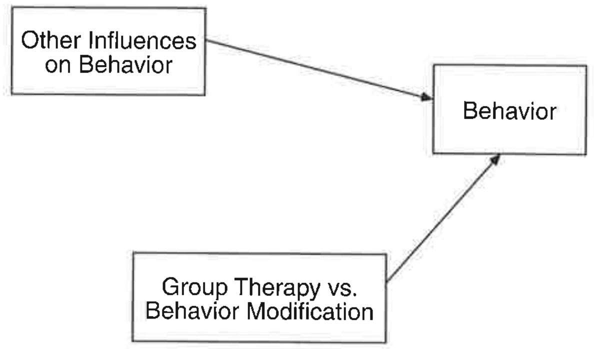

Figure 12.6 A true experiment in path form. Due to random assignment to groups (Group Therapy versus Behavior Modification), the variables that affect Behavior (the effect) do not affect the cause (treatment group). Random assignment rules out common causes, which is why we don't need to control for the multitude of other influences on Behavior in our analysis.

had only a minor effect on this estimate, which changed from .160 to .165. We could exclude Ethnicity from this model without seriously affecting the estimate of the influence of Parent Involvement on GPA. To reiterate, models must include common causes of the presumed cause and the presumed effect if they are to be valid, but they need not include *all* causes.[2]

True Experiments and Common Causes

The elimination of the danger of omitted common causes is the reason that true experiments allow such a powerful inference of cause and effect. As a general rule, experimental research, in which participants are *assigned at random* to experimental versus control groups, has a higher degree of internal validity than does nonexperimental research, meaning that it is generally less dangerous to make an inference of cause and effect with a true experiment. Figure 12.6 helps illustrate the reason for this power. Suppose you conduct an experiment in which you assign, at random, children with behavior disorders to two types of treatments: group therapy or behavior modification, with some measure of behavior improvement as the dependent variable. Figure 12.6 illustrates this experiment in path analytic form, with the path from the dummy variable Group Therapy versus Behavior Modification to Behavior providing the estimate of the relative effectiveness of the two treatments. But a multitude of variables affect children's behavior, from parents to friends to teachers, and many more. Why don't we have to consider these variables when we conduct our analysis of the experiment? The reason we don't have to consider these Other Influences on Behavior is because they are not *common* causes of assignment to treatment groups and Behavior. Although these Other Influences affect Behavior, they did not affect assignment to the Therapy versus Behavior Modification groups because assignment to the treatment groups was random, based on the flip of a coin. This, then, is why true experiments are so powerful. True experiments still require an inference of cause and effect, but we can make that inference so powerfully because the act of random assignment effectively excludes all possible common causes of the presumed cause and the presumed effect. Random assignment assures that no other variables affect the presumed cause.

Intervening (Mediating) Variables

Given the admonition that models must include all common causes of the presumed cause and presumed effect, you may wonder how this applies to intervening or mediating variables.

Do you also need to include all variables that *mediate* the effect of one variable on another? The answer is no; mediating variables are interesting because they help explain how an effect comes about, but they are not necessary for the model to be valid; in short, they are gravy. It is good that mediating variables are not required to make the model valid, because you could always include another layer of mediating variables. In the present example, you might wonder if Homework and TV viewing time mediate the effects of Parent Involvement on GPA (cf. Keith et al., 1993). That is, do parents influence their adolescents' learning, in part, by influencing them to complete more homework and watch less TV? Suppose you found that these variables indeed mediated the influence of Parent Involvement on GPA; you might then wonder if the effects of Homework were mediated by time on task, and so on. Even with a seemingly direct relation, say the effect of smoking on lung cancer, we could posit and test indirect effects—the effect of smoking on buildup of carcinogens in the lungs, the effect of these chemicals on the individual cells, and so on. Again, it is not necessary to include indirect effects for models to be valid, but such indirect effects can help you understand how effects come about.

In our current example, suppose that our central interest was the effect of Previous Achievement on GPA. If we were to conduct an analysis examining the direct effect of Previous Achievement on GPA without the intervening variable of Parent Involvement, the standardized direct (and total effect) would be .472. If you conduct the calculations for the indirect and total effects, you will find that the total effect of Previous Achievement on GPA for Figure 12.3 is also .472. When mediating or intervening variables are included in the model, the total effects do not change (although direct effects do); indirect effects are unnecessary for model validity.

I stress again, however, that although unnecessary for valid models, indirect effects are often very illuminating. Our current example suggests that Parent Involvement has a positive effect on GPA. But how does that effect come about? Previous research that tested for possible mediation by homework and TV viewing suggests that homework, in fact, partially mediates the effect of parent involvement on learning but that TV viewing does not (Keith et al., 1993). Parents who are more involved encourage, cajole, or force their children to do more homework, and this homework, in turn, raises their achievement. Although parents who are involved also influence their adolescents to spend less time watching TV, TV viewing appears to have little effect on achievement. Thus, leisure TV viewing does not appear to mediate the effect of parent involvement on achievement. As you become more expert in a particular area of research, you will likely find yourself asking questions about indirect or mediating effects. Indeed, even for those conducting experiments, indirect effects may often be of interest. Suppose you find that your experimental treatment (e.g., a new versus an established type of consultation) is effective; you may next reasonably wonder why. Is it because the new consultation method improved problem identification, or speeded the time to intervention, or made evaluation more complete? Another advantage of mediating variables is that they can help strengthen the causal inferences embedded in path models. Logically, if you can explain both which variables affect an outcome and the mechanism *by which that effect occurs*, your causal claims are more believable. If we can demonstrate the indirect effect of smoking on lung cancer through the buildup of carcinogens in the lungs, it strengthens the case for smoking, as opposed to other characteristics of smokers, being a cause of lung cancer (Pearl, 2009). For additional information on testing mediating variables, see Baron and Kenny (1986), MacKinnon (2008); MacKinnon et al. (2002), or Shrout and Bolger (2002; see also the earlier discussion of mediation in Chapter 8.

OTHER POSSIBLE DANGERS

Paths in the Wrong Direction

Another possible danger in path analysis (and nonexperimental research in general) is that you may draw a path in the wrong direction. The implications of this danger depend on where this mistake takes place.

Figure 12.7 shows a model in which I erroneously assumed that 10th-grade GPA affected 8th-grade Parent Involvement. This model is clearly impossible, because it violates one of our primary assumptions, that cause cannot happen backward in time. The GPA variable occurs in 10th grade (although it is actually a measure of 9th- and 10th-grade GPA), whereas the Parent Involvement variable occurs in 8th grade. The model is clearly impossible. There is, however, nothing in the multiple regression analyses and nothing in the figure that would alert you to the fact that your model is incorrect. Indeed, the model leads you to completely erroneous conclusions about the moderate effect of 10th-grade GPA on 8th-grade Parent Involvement. Obviously, if the arrow between the two variables of prime interest is drawn in the wrong direction, the results will be completely and totally misleading.

In contrast, Figure 12.8 shows a model in which the path between Previous Achievement and Parent Involvement is drawn in the wrong direction. Previous Achievement was included in the model as a potential common cause of Parent Involvement and GPA, so the model in Figure 12.8 no longer controls for this variable as a *common* cause. Again, there is nothing in any of the analyses to suggest that our model is incorrect. In this case, however, with the mistaken path being between our primary causal variable and a "control" variable, the findings are not quite as misleading. In fact, the direct effects of each variable in the model are the same as they were with the "correct" model shown in Figure 12.3. This makes sense when you realize that all paths to GPA are estimated via simultaneous MR, and for the models shown in Figures 12.3 and 12.8, both simultaneous regressions regressed GPA on each of the four variables in the model. What are incorrect in Figure 12.8 are the *total* effects. In Figure 12.3, Previous Achievement has an indirect effect on GPA through Parent Involvement, and thus

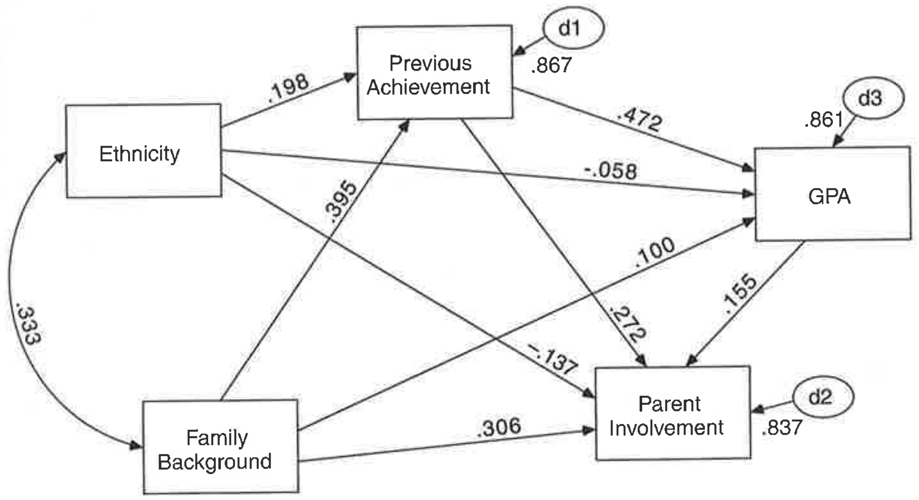

Figure 12.7 In this model the path between GPA and Parent Involvement is drawn in the wrong direction. There is nothing in the results to indicate that it is wrong.

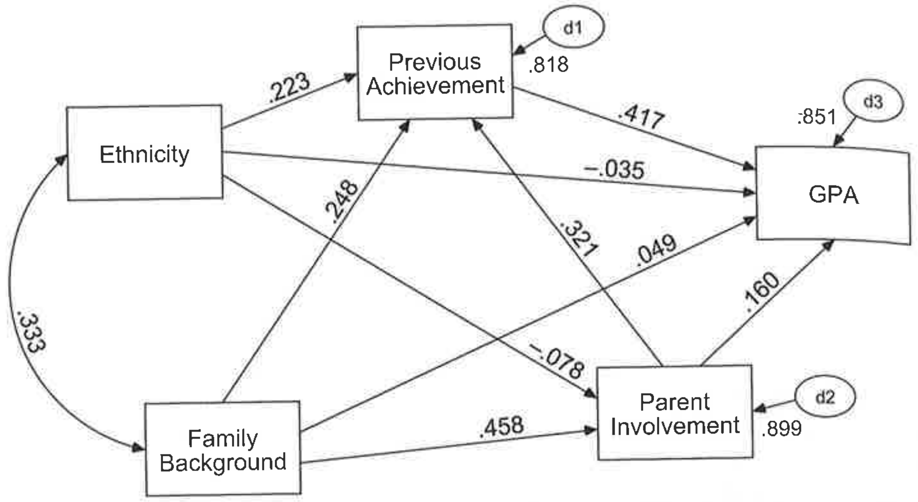

Figure 12.8 In this model the path between Parent Involvement and Previous Achievement is drawn in the wrong direction. The direct effects for these two variables remain the same, but the indirect and total effects differ from the "correct" model in Figure 12.3.

its total effect is .472, compared to a direct effect of .417. In Figure 12.8, Previous Achievement is the next to last variable in the causal chain, so it has no indirect effect on GPA; its direct and total effects are both .417. For Parent Involvement, the reverse is true. In the "correct" model (Figure 12.3), Involvement had no indirect effect on GPA, so its direct and total effects were both equal to .160. In Figure 12.8, Involvement has an indirect effect on GPA through Previous Achievement, and thus we overestimate its total effect as .294. You should calculate the indirect and total effects yourself to make sure your estimates agree with mine, and that you understand the difference between these two models.

If the variables with paths drawn in the wrong direction are two of the less central variables, there should be little or no effect on the estimates of the primary variables of interest. For example, suppose the current example included a path from Ethnicity to Family Background, rather than a correlation. Suppose further that we erred by drawing that path in the wrong direction (from Family Background to Ethnicity). This mistake will have no effect on the estimates of the direct, indirect, or total effects of Parent Involvement on GPA.

To summarize, if the effect, the final endogenous variable, is in the wrong position, estimates of all effects will be erroneous. If the primary causal variable has paths drawn in the wrong direction (but not the primary *effect* of interest), estimates of direct effects may still be accurate, but indirect and total effects will likely be incorrect. If background variables have paths drawn in the wrong direction, this error will likely not affect estimates of effects from the main cause variable to the main effect. These comments apply to just-identified models estimated through multiple regression but are not too far off for other, more complex models.

Reciprocal Causal Relations?

Given the problems resulting from paths drawn in the wrong direction, you may be tempted to be open-minded and proclaim that the variables are causally related in a reciprocal fashion in which not only does *a* affect *b* but *b* also affects *a*. Don't succumb to this temptation at this stage of your development! Although it is indeed possible to estimate such nonrecursive models, you cannot do so using multiple regression. You can estimate nonrecursive models using the SEM programs discussed in subsequent chapters, but such models are neither easy nor their results always illuminating. In my experience, reciprocal effects are also less common than you might think. Reserve the use of nonrecursive models for those cases in which you really think reciprocal effects may exist or for which you have legitimate, substantive questions about causal direction, not those for which you are simply unsure.

An even worse solution to this dilemma is to try to conduct the regression–path analysis both ways to see which "works best." You have already seen that the results of simple path analyses do not tell you when you have a path in the wrong direction. Likewise, the results of the analyses do not inform you as to which direction is best. Once again, theory, previous research, and logic are the appropriate tools for making such judgments.

I should note that, although the results of just-identified path analyses estimated through multiple regression cannot inform decisions about causal direction, properly overidentified models estimated through an SEM program may indeed be able to help with such decisions. In addition, well thought out nonrecursive models estimated via SEM programs can also be very informative about the nature and process of how one variable affects another. We will discuss these issues in later chapters.

Unreliability and Invalidity

One assumption underlying the causal interpretation of regression and path coefficients is that the exogenous variables are measured with near perfect reliability and validity. With our current model, Ethnicity may come close to meeting this assumption, but the variable Family Background, a composite of Parent Education, Parent Occupational Status, and Family Income, certainly does not. We obviously regularly violate this assumption but will postpone until later chapters a discussion of the effects of this violation and possible solutions.

DEALING WITH DANGER

The two primary dangers of path analysis are (1) that you have neglected to include in your model an important common cause of the variable you think of as your primary cause and the variable you think of as your primary effect and (2) that you have drawn paths in the wrong direction; that is, you have confused cause and effect. In the jargon of SEM, these are generally termed specification errors, or errors in the model. Of these two, I consider the first the most common and insidious. In most cases, it should be pretty obvious when you draw a path in the wrong direction. What can you do to avoid these errors?

My first response is to say, "Welcome to the dangerous world of structural equation modeling; join us SEMers on the wild side!" More seriously, I again remind you that these same dangers apply to *any* nonexperimental research, no matter how that research is analyzed. One advantage of path analysis and structural equation modeling, in my opinion, is the requirement of a theory, generally expressed figurally in a path model, prior to analysis. It is much easier to spot missing common causes and causal assumptions in the wrong direction when examining a path model than it is when reading the description of, say, a MR analysis. Furthermore, these dangers apply to *all* research, experimental or nonexperimental, in which we wish to infer cause and effect. A true experiment allows a powerful inference of cause and effect by knocking one leg out from under the danger of common causes, but the farther we stray from the true experimental ideal of random assignment to treatment groups, the more real this danger becomes. Indeed, many concerns with quasi experimental research (e.g., research using matched groups rather than random assignment) boil down to concerns over unmeasured common causes. With a true experiment we also actively manipulate the independent variable, the presumed cause, thus making true experiments less likely to confuse causal direction, as well.

We have seen that the analyses themselves do not guard against these errors; they do not tell us when our models are wrong or when we have neglected an important common cause. How, then, to avoid these specification errors? I come back to the same refrain: understand relevant theory; be familiar with the research literature; spend time puzzling over your model, especially thinking about potential common causes and potential problems in direction; and draw your model carefully.

These same concerns and dangers apply when you are a consumer and reader of others' research. As you read others' nonexperimental research, you should ask yourself whether the researchers neglected to include any important common causes of their presumed cause and presumed effect. If so, the results of the research will be misleading and likely overestimate (or underestimate) the effect of one variable on another. Arm-chair analysis is not sufficient, however; it is not valid to simply say, "Well, I think variable *Z* is a probable common cause of variables *X* and *Y*," and expect to have your concerns taken seriously. You should be able to demonstrate, through theory, previous research, or analysis, that variable *Z* is indeed a likely and important common cause. Likewise, as you read nonexperimental research, you should be attuned to whether any of the causal assumptions are reversed. Again, you should be able to demonstrate this incorrect causal direction through theory, research, logic, or your own analyses.

We will revisit the danger of measurement error and its effects. For the time being, you should simply strive to make sure that all your variables, and especially your exogenous variables, are as reliable and valid as possible.

REVIEW: STEPS IN A PATH ANALYSIS

Let's review the steps involved in path analysis now that we've carefully considered the dangers.

1. First, spend some time thinking about the problem; how might these variables of interest be causally related?
2. Draw a tentative model.
3. Study relevant theory and research. Which variables must be included in the analysis? You must include the relevant common causes, but not every variable under the sun. The relevant theory and research, along with careful thought, will also help you resolve questions of causal direction. "The study of structural equation models can be divided into two parts: the easy part and the hard part" (Duncan, 1975, p. 149). This step is the hard part of SEM.
4. Revise the model. It should be lean, but include all necessary variables.
5. Collect a sample and measure the variables in the model, or find a data set in which the variables are already measured. Use reliable and valid instruments.
6. Check the identification status of the model. Make sure the model is just-identified or overidentified.
7. Estimate the model.
8. Fill in the model estimates (paths and disturbances) in your figure. Are the paths more or less as expected? That is, are the paths you expected to be positive in fact positive; those you expected to be negative, negative, and those that you expected to be close to zero in fact close to zero? Meeting such expectations allows more confidence in your model.
9. Write up the results and publish them.

Some writers recommend *theory trimming* in between my steps 8 and 9. Theory trimming means deleting statistically nonsignificant paths and re-estimating the model. I do not recommend this step, especially when using multiple regression to solve for the paths. We will return to this issue in the next chapter.

SUMMARY

The chapter began by reiterating the basic assumptions of multiple regression: linearity, independence, and homoscedasticity. For regression coefficients to provide accurate estimates of effects, the disturbances should be uncorrelated with the exogenous variables. This assumption will likely be fulfilled if there is no reverse causation, the exogenous variables are perfectly measured, equilibrium has been achieved, and there are no omitted common causes in the model.

These assumptions led to a discussion of the dangers of path analysis. When a common cause (a variable that affects both a presumed cause and a presumed effect) is omitted from a model, this omission changes the estimate of the influence of one variable on another. The most common result is that we end up overestimating the effect, although underestimation is also possible. The dreaded *spurious correlation* is a result of an omitted common cause, and thus omitted common causes are the primary reason for the admonition about inferring causation from correlations. I illustrated the effects of omitting a common cause through a research example testing the effects of Parent Involvement on 10th-grade GPA. When Previous Achievement, a common cause of Involvement and GPA, was omitted from the model, we overestimated the effect of Parent Involvement on GPA. Omitted common causes may be a reason for variability in research findings in nonexperimental research.

The warning to include common causes should not be interpreted as a mandate to include all causes of the presumed cause and the presumed effect. Only variables that affect both the presumed cause and presumed effect must be included. We illustrated the difference between a cause and a common cause by deleting Ethnicity from the model. Ethnicity affected Parent Involvement but not GPA, and thus was not a *common cause* of the two variables. As a result, when Ethnicity was removed from the model, the estimate of the effect of Involvement on GPA barely changed. The main reason that true experiments allow such a powerful inference of causality is because, through the act of random assignment, such research rules out possible common causes of the independent (cause) and dependent (effect) variable, even though experiments do not rule out all causes of the dependent variable.

The warning to include common causes also does not extend to mediating or intervening variables. When an intervening variable is included in the model, the total effects remain the same, but a portion of the direct effect of *X* on *Y* becomes indirect effect through the mediating variable. Intervening variables help explain *how* an effect comes about but do not need to be included for the model to be valid.

Estimates of effects are also incorrect when paths are drawn in the wrong direction, although the extent of the problem depends on the paths involved. If the incorrect path is from the effect to the cause, the results will obviously be incorrect and completely misleading. If the incorrect path involves the primary causal variable and one of the other causal variables in the model, this error will affect the total effects but not the direct effects. If the incorrectly drawn path involves some of the background variables in the model, this error should have little effect on the estimates of primary interest (although it will make attentive readers less trusting of your results!). We will revisit and address this danger in subsequent chapters.

How, then, can you be sure that your model is correct? Have a good understanding of relevant theory and previous research. Think about the variables in your model, how they are related to one another. If necessary, bolster causal assumptions (e.g., *a* affects *b*, rather than *b* affects *a*) through the use of longitudinal data. Think about possible common causes and investigate them in the research literature. If necessary, test common causes in the research itself. In fact, most of what you should do to ensure the adequacy of your model boils down to the same advice for drawing a model in the first place: theory, previous research, and logic.

I also noted that, as a reader or reviewer of others' nonexperimental research, it is not enough to guess about neglected common causes; you should be able to demonstrate such criticisms through theory, previous research, or independent analysis. Finally, I noted again that these dangers apply to all nonexperimental research, no matter how it is analyzed. One advantage of path models is that the figural display of one's model (in essence a mini theory) often makes errors and assumptions more obvious and therefore more likely to be corrected. We postponed dealing with the violation of the assumption of perfect measurement until later chapters.

EXERCISES

1. Conduct each of the parent involvement analyses reported in this chapter, using the NELS data. The variables, as listed in NELS, are: Ethnicity = Ethnic; Family Background = BySES; Previous Achievement = ByTests; Parent Involvement = Par_Inv; and GPA = FfuGrad. Compare your results to mine.
   a. Make sure you understand what happens when a common cause is omitted versus a simple cause of only one of the variables of interest (Figures 12.3 through 12.4). Is Family Background a common cause or a simple cause of Parent Involvement and GPA? Try deleting it from the model; what happens to the path from Involvement to GPA?
   b. Analyze a model without Parent Involvement. Calculate direct, total, and indirect effects for each variable on GPA. Do the same for the model shown in Figure 12.3. Compare the tables of direct, indirect, and total effects.
   c. Analyze a model like Figure 12.3, but in which a path is drawn from Ethnicity to Family Background. Now analyze a model in which the path is drawn from Family Background to Ethnicity. Which model is correct? How did you make this decision? What effect, if any, did this change in direction have on the estimate of the effect of Parent Involvement on GPA?
2. Find an article that uses path analysis or explanatory multiple regression on a research topic with which you are familiar and interested. If the authors' model is not drawn in the article, see if you can draw it from their description. How do the authors justify their causal assumptions or their paths? Do you agree, or do you think some of the paths are drawn in the wrong direction? Do you think there are any obvious common causes that have not been included in the model? Can you demonstrate that there are common causes that have been neglected? If the authors included a correlation matrix with their article, see if you can reproduce their results. Draw the estimated model.
3. In Chapter 11, you constructed and tested a path model using the variables Family Background (BYSES), 8th-grade GPA (BYGrads), 10th-grade Self-Esteem (F1Concpt2), 10th-grade Locus of Control (F1Locus2), and 10th-Grade Social Studies Achievement (F1TxHStd). Refer to or redo the analysis. For the sake of consistency, make sure you have Social Studies Achievement as the final endogenous variable.
   a. Notice the direct effects of Self-Esteem and Locus of Control on Social Studies Achievement. Focus on the effect of GPA on Self-Esteem and Locus of Control. Is 8th-grade GPA a common cause of these variables and Social Studies Achievement? Now remove the 8th-grade GPA variable from the model. What happens to direct effects of Self-Esteem and Locus of Control on Social Studies Achievement? Explain the difference in effects from the original model.
   b. Did you draw a path from Self-Esteem to Locus of Control or Locus of Control to Self-Esteem? Calculate the direct, indirect, and total effects of these two variables on Social Studies Achievement. Whichever way you drew the path, now reverse the direction and re-estimate the model. Recalculate the direct, indirect, and total effects of these variables on Social Studies. Explain the differences you find.

Notes

1 If the common cause has positive effects on both the presumed cause and the presumed effect, its neglect will lead to an overestimate of the effect of the presumed cause on the presumed effect. If a common cause has a negative effect on either variable, its omission will lead to an underestimate, and if it has a negative effect on both, its omission will result in an overestimate.

2 There may be other advantages for including a variable in the model that is not a common cause. For example, inclusion of noncommon causes results in overidentified models, the advantages of which we will discuss in the following chapter.