Statistical Methods for Communication Science

ANDREW F. HAYES
The Ohio State University

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2005 Mahwah, New Jersey London

CHAPTER

SIXTEEN

Interaction

Up to this point, we have discussed how to quantify and test for association between variables, determing whether a relationship exists between two variables after accounting for their shared relationship with another variable or set of variables, and how to determine whether "chance" is the most parsimonious explanation for differences between groups or a relationship that we observe between two variables. But there are questions that need to be answered before we can say we really have gained some understanding of communication processes and theory from our research: "Why?" "When?" and "How?" A scientist can make a career of demonstrating that two variables are related, but the more memorable studies, the more impressive studies, and ultimately the more influential studies in the field go further by discovering or explaining why such relationships exist, under what circumstances, or for whom the relationship exists strongly as opposed to weakly or not at all. We truly understand some phenomenon if we are able to determine when the phenomenon will occur, why or how it occurs, and for whom it occurs or will occur. As you become increasingly knowledgeable about the discipline of communication and increasingly expert in your specialty area, you will discover that the most sensible answer to almost every question you will confront as a scientist is "it depends." Such an answer is not a cop-out. What we study is often sufficiently complicated that it would be incorrect to say without condition or exception that X causes Y or that one group differs from another group on some outcome variable in all circumstances. Usually effects vary as a function of something else. For example, perhaps for some people exposure to televised violence causes aggressive behavior, but for others such exposure has no effect. Or perhaps the effect of such exposure differs depending on the consequences or form of the violence. Media violence that is perceived to be rewarded may lead people to engage in that behavior, whereas violence that is perceived to be punished may discourage such behavior. Or perhaps a message about the negative consequences of unsafe sexual practices could increase safe sexual practices among people of a certain background or age but decrease it or have no effect among people of a different background or age.

If a relationship between X and Y varies depending on the value of some other variable W, then it is said that W is a *moderator* of the relationship between X and Y, or that the relationship between X and Y is *moderated by* W. In other words, W moderates the relationship between X and Y if the value of W predicts the size or direction of the relationship between X and Y. Another term often used to describe moderation is *interaction.* We say that two variables X and W interact if the combination of X and W explain variation in Y independent of their additive effects. Thus, interaction is akin to the concept of synergy—when two things are combined they have a different effect than the sum of their parts. I will use the terms *interaction* and *moderation* interchangeably throughout this chapter.

The concept of interaction is perhaps more easily understood with a picture. Figure 16.1 illustrates three forms of interaction, as well as 3 examples of the absence of interaction. In the top row on the left, the relationship between X and Y, expressed as a regression line, varies depending on the value of W, where W can have only two values (e.g., $W = 0$ for males and $W = 1$ for females). But a lack of interaction between X and W is displayed in the top right panel. It is clear in that graph that the relationship between X and Y does not vary across the two groups defined by W, reflected in the fact that the slope of the regression line estimating Y from X is the same for both values of W. But W need not be dichotomous, as the middle two panels indicates. The graph in the left panel, middle row, depicts a relationship between Y and X that varies as a function of the values of W, whereas the relationship does not differ as a function of W in the right middle panel. The bottom row left panel illustrates how the relationship between X and Y might differ as a function of whether participants are assigned to an experimental or a control condition in an experiment. So the relationship between X and Y depends on the level of the experimental manipulation a participant was assigned to. A corresponding lack of interaction is displayed in the right panel of the bottom row. It should be apparent from these examples that interaction or moderation evidences itself graphically in the form of nonparallel regression lines. In all these examples on the left, the effect of X on Y (represented with the regression line) depends on some value W. A lack of interaction shows up graphically as parallel regression lines, reflecting the fact that the relationship between X on Y remains constant across all values of some third variable W.

In this chapter, I introduce some statistical approaches to testing for interaction between two predictor variables. This may be the most complicated of all chapters in the book, but it is arguably one of the more important chapters, because moderation is such a commonly tested hypothesis in communication science. It is also one of the more incomplete chapters in the book. We only begin to scratch the surface of statistical approaches to testing for interaction and the various forms that interaction can take. Whole books have been written on this topic (e.g., Aiken & West, 1991; Aquinis, 2002; Jaccard, Turrisi, & Wan, 1990), and there are literally dozens upon dozens of articles in the methodology literature about statistical interaction. But before focusing on the nuts and bolts of testing for interaction, let's first look at some examples in communication theory and research.

16.1 Interaction in Communication Research and Theory

Many of the hypotheses communication researchers test focus on interaction or moderation, and many of the theories that explain communication phenomena involve interaction between components of the theory. For example, Walther, Slovacek, and Tidwell (2001) were interested in how nonverbal information such as information contained in a person's face might affect the relational outcomes of a computer-mediated communication (CMC) task, and whether such an effect depends on whether the CMC

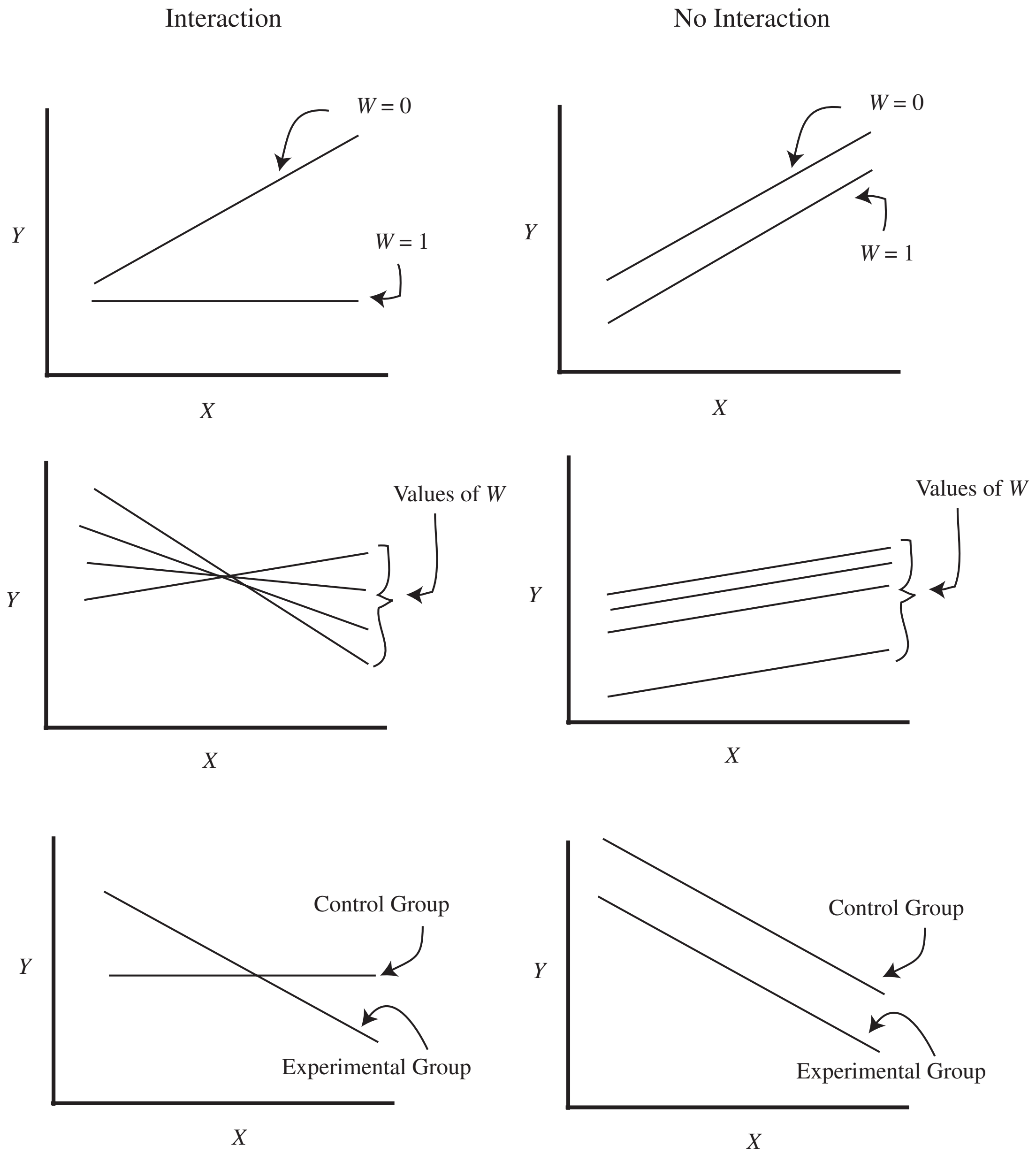


Figure 16.1 Graphical representation of interaction and lack of interaction.

partners had known each other for a long or relatively short period of time. In this study, Walther et al. (2001) manipulated whether the participants were given a photo of their interaction partners in a CMC context. They were also able to categorize the participants into two groups: whether the partners had interacted with each other in a CMC context only briefly during the procedure or had been interacting with each other more extensively over a long period of time. They found that the effect of the photograph on later judgments of feelings of intimacy and attraction toward the partners differed systematically as a function of the length of the CMC relationship. More specifically, they reported that in short term CMC relationships, a visual image of the interaction partners enhanced feelings of intimacy and attraction toward those partners, but in long-term CMC relationships, the presence of the photograph reduced such feelings. So we can say that the effect of the photograph on feelings of intimacy was moderated by the length of the CMC relationship or that the presence or absence of the facial information interacted with the length of the relationship in explaining variation in perceived intimacy.

Another example comes from research and theory on the knowledge gap (Tichenor, Donohue, & Olien, 1970). The knowledge gap refers to the differences in knowledge possessed by the "haves" and "have-nots" in society and the differential effect of information on such groups. More specifically, people who are relatively low in socioeconomic status tend to have less knowledge about a number of things, such as politics or world affairs, than people higher in socioeconomic status. There are a number of explanations for this phenomenon. One explanation is that people who are lower in education tend to have fewer of the cognitive skills and less of the background knowledge to make sense of information presented through the mass media. Thus, increased exposure to mass-mediated information is less likely to facilitate learning among the relatively less educated. But greater exposure should enhance learning among the more educated because they have the skills and aptitudes and prior knowledge that would help them to better understand and therefore learn from the mass media. Indeed, in a study to test this possibility, this is exactly what Eveland and Scheufele (2000) found. Using data from the 1996 National Election Study, they found that individual differences in media use were more predictive of individual differences in political knowledge among the more educated than among the less educated. Thus, education moderated the relationship between news exposure and political knowledge. That is, education and news exposure interacted in explaining person-to-person variation in political knowledge.

Yet another example is found in cultivation theory and the notion of *mainstreaming*. Cultivation theory attempts to explain the effects of television on the beliefs and attitudes of the public. According to cultivation theory, greater exposure to television leads to a greater internalization of the "television view" of the world. Thus, the more television a person watches, the more likely his or her perceptions of the world and the attitudes he or she holds will come to mirror the stories, world views, and attitudes that predominate the televised world. The notion of mainstreaming refers to the homogenizing effect that exposure to television produces among heavy viewers. People of different ethnicities, levels of education, or political orientations often have very different attitudes about social issues and different beliefs about the world, such as how dangerous it is or how trustworthy people are. The mainstreaming hypothesis predicts that individual differences such as ethnicity or education should be less related to attitudes or perceptions of the world among heavy viewers of television compared to light viewers, because heavy doses of the televised world leads to a convergence of the beliefs, attitudes, and perceptions of otherwise disparate groups of people toward the televised view of the world. Among light viewers, those who are less likely to have experienced the cultivating effect of television, such individual differences as education, ethnicity, and political orientation are more predictive of beliefs, attitudes, and perceptions because light viewers' world views are less likely to have been shaped by the images of television. Thus, cultivation theory and the mainstreaming hypothesis argue for a moderating effect of television viewing on the relationship between demographic variables such as education, ethnicity, and political orientation and beliefs, attitudes, and/or world views. In other words, frequency of television and such demographics are proposed to interact in the explanation of individual differences in beliefs, attitudes, or perceptions of the world. Although cultivation theory remains controversial, there is at least some evidence in the communication literature supporting cultivation theory and the mainstreaming hypothesis, summarized in such places as Gerbner, Gross, Morgan, and Signorielli (1986), and Shanahan and Morgan (1999).

Finally, the elaboration likelihood model of persuasion (Petty & Cacioppo, 1986) predicts an interaction between characteristics of a message or its source and a person's motivation or ability to process that message in determining how persuaded a person will be by that message. According to the elaboration likelihood model, people who are more motivated or more able to engage in thoughtful processing of message content are likely to be influenced by such features of a message as the strength of the arguments contained within it. People with relatively little motivation or ability, in contrast, are more influenced by the presence of "peripheral cues" of a message such as whether the source is likeable or attractive or the sheer number of arguments presented rather than their quality. Decades of research (summarized in Petty & Cacioppo, 1986) illustrates such interactions. For example, the effect of argument quality on persuasion depends on whether the content of the message is relevant to a person's life. Such "personal involvement" in an issue leads to deeper processing of message content, such that messages with predominantly strong arguments induce greater attitude change and greater memory for the content of the message than do messages with predominantly weak arguments. But when personal involvement is low, people are less likely to evaluate a message in terms of the quality of the arguments because they are less likely to engage in the kind of thoughtful message processing that would be required to determine whether a set of arguments is strong or weak. So attitude change and memory for arguments is largely unrelated to whether the arguments in the message are strong versus weak.

These examples all illustrate that much communication research and theory is based on the notion of interaction. Variables do not have consistent effects in the communication literature, or communication theory predicts that a variable's affect on some outcome will vary as a function of some other variable. For one reason or another, a variable may have one type of effect in some circumstances or among some people, but have a different effect in some other circumstance or among some other group. The ability to test hypotheses that focus on differences in effect and to discover such differences if they exist is an important skill that the communication researcher must possess.

There are two primary contexts in which questions about moderation are usually addressed statistically, with those contexts being defined in terms of whether all or only some of the predictor variables proposed to be interacting are categorical. When the predictor variables are all categorical, interaction is usually tested with *factorial analysis of variance*. For example, variables X and W may be experimental manipulations and the researcher is interested in knowing if the manipulation of variable X has the same effect across the levels of the manipulation of W. Or W may be a naturally occurring categorical variable like gender, ethnic group, or any other conceivable nominal variable. In that case, the question focuses on whether the experimental manipulation has the same effect in all groups defined by variable W. But if one or both of the predictor variables are quantitative dimensions, *moderated multiple regression* is more appropriate. In moderated multiple regression, the question focuses on whether the regression weight estimating Y from X varies as a function of some second variable W. Variable W can be either nominal with two or more possible categories, or quantitative with many possible values.

If W is categorical but X is quantitative, it is all too common to categorize cases in the data file based on their scores on X and then apply factorial analysis of variance rather than use the more efficient method of moderated multiple regression. I strongly discourage this strategy for reasons discussed toward the end of this chapter. This practice is common perhaps because factorial analysis of variance is a bit easier to grasp and therefore is probably more widely taught, understood, and therefore used. For this reason, I focus first on factorial analysis of variance. But as you will see, factorial analysis of variance is just a special form of multiple regression.

16.2 Factorial Analysis of Variance

Berger (2000) was interested in how media reports of increasing crime in a community contribute to people's perceptions of their risk of being a victim of crime. The media often reports frequency information about crime in a community over time and uses upward trends as evidence that crime is increasing. For example, a daily newspaper might report that in 2000 there were 200 crimes in Anytown whereas in 2005, there were 250 crimes. Understandably, knowledge that crime in your community has increased could make you feel uneasy and vulnerable. But such an increase in the frequency of crime would not be at all surprising if the population of Anytown also increased between 2000 and 2005. The more people there are in a region, the more crimes there are going to be, because there are more people, houses, business, etc. Berger (2000) argued that if the media included in their stories such information about population growth trends along with information about the trends in the frequency of crime, then the effect of information about upward trends in crime over time on people's feelings of vulnerability would be reduced. But he argued that such population trend information would not affect everyone the same. Specifically, he hypothesized that the reduction in perceived risk of being a victim of crime associated with the additional information about the size of the population over time would be smaller for women. This prediction was based on previous research that women seem to feel more vulnerable to crime than men and that this elevated feeling of vulnerability would interfere with a woman's ability to connect the information about the increase in population to the increase in the total number of crimes.

To test this hypothesis, a group of men and a group of women read a short news article describing how there had been an increase in the number of burglaries in the community in which they lived over a 5 year period. Half of the participants randomly assigned to the *Information Present* condition also read a second news article that described how the size of the population during the period had increased during this same 5-year period. The other half of the participants, randomly assigned to the *Information Absent* condition, did not get this story. After reading the story (or stories), the participants were asked a series of questions, including one that asked them to rate the likelihood that they would be a victim of a burglary on a 0 (certainly not) to 100 (certainly) scale. This was the dependent variable in their analysis that we will call *risk* or *perceived vulnerability.*

Berger (2000) was hypothesizing an *interaction* between gender and whether or not the participant received the population trend information on the participants' perceived risk judgments. That is, he proposed that the difference in perceived risk between men who received the population trend information and those who did not should be smaller than the corresponding difference in women. Rephrased, the size of the effect of population trend information on risk judgments should depend on whether the reader was a male or a female. If we call μ mean risk judgment, then the null and alternative hypothesis are

$$H_0 : (\mu_{MA} - \mu_{MP}) = (\mu_{FA} - \mu_{FP})$$
$$H_a : (\mu_{MA} - \mu_{MP}) \neq (\mu_{FA} - \mu_{FP})$$

where the first subscript refers to sex (**M**ale or **F**emale) and the second subscript refers to the population trend information (**P**resent or **A**bsent). The first difference in parentheses, $\mu_{MA} - \mu_{MP}$, is the effect of population trend information on the risk judgments of men, whereas the second difference, $\mu_{FA} - \mu_{FP}$ is the effect of population trend information on women. So the null hypothesis states that there is no difference between men and women in the effect of population trend information, whereas the alternative states that the effect of population trend information differs between men and women. But notice that as the research hypothesis is phrased, a one-tailed test is justified. That is, the research hypothesis could be framed statistically as

$$H_a : (\mu_{MA} - \mu_{MP}) < (\mu_{FA} - \mu_{FP})$$

but we will stick with two-tailed tests here because in more complicated ANOVA designs, it often isn't possible to test a directional alternative because of the way that ANOVA works mathematically. Furthermore, it is sensible to remain open to the possibility that the result could be the opposite of what was predicted.

Before continuing, a comment about the use of Greek symbols in the null and alternative hypothesis is warranted. As you know by now, is conventional in statistics to use Greek letters to refer to characteristics of a population and Roman letters to refer to characteristics of a sample from that population. The Greek letter μ is typically used to denote a population mean and a Roman character such as $\overline{Y}$ to refer to the mean computed of a sample from some population. In experimental contexts, the notion of a population is a bit different than in nonexperimental studies. In nonexperimental studies, the population refers to the universe of units (e.g., people) from which the sample was derived. In experiments, we often use the term "population" to refer to something more hypothetical. Consider μ_{MA}. This notation refers to a hypothetical population of males and what their average risk judgment would be expected to be in the information absent condition of the study. Of course we don't know μ_{MA}. At best, we can estimate this by obtaining some men and putting them in this condition and seeing what their judgments are. The population is strictly hypothetical because the experimental context is a world that we are creating. It doesn't exist in reality. There is no population of men who read stories about crime without corresponding information about population change over time. But imagine if we had unlimited resources and could conduct this study with a very large number men. If we could do this, then the sample mean, $\overline{Y}_{MA}$ would probably be a pretty good descriptor of how men, when placed in the *Information Absent* condition, would be expected to respond when asked how vulnerable they feel to burglary. Similarly, $\overline{Y}_{MP}$ would be a pretty good descriptor of how men, when placed in the *Information Present* condition, would be expected to respond. If the population trend information information has no effect on risk judgments in men, we are in making the claim that $\mu_{MA} = \mu_{MP}$ or, equivalently, $\mu_{MA} - \mu_{MP} = 0$. The same logic applies to the population means for women.

To test the hypothesis of interaction, the standard statistical method used is *factorial analysis of variance*. In analysis of variance, the independent variable or variables are often called *factors*, and the values of each factor are referred to as *levels*. In this study, there are two independent variables or factors, defined as gender and population trend information, each with two levels (male vs. female and population trend information present vs. absent). Thus, the analysis strategy described here is a "2 × 2" (pronounced "two by two") *between-groups factorial analysis of variance*, with the "2" referring to the number of levels of the factors. The "factorial" label comes from the fact that these two factors are perfectly crossed with each other, such that each each level of one factor occurs in the design at each level of the second factor. The factors in a factorial ANOVA can have any number of levels, but we will only discuss the 2 × 2 case in this chapter. The "between-groups" part of this description refers to the fact that each participant contributes data to one and only one of the 4 *cells* in this design. Each cell is defined by the combination of levels of the factors. So the four cells in the design are (a) males, information present, (b) males, information absent, (d) females, information present, and (d) females, information absent. Other types of factorial ANOVA commonly conducted in communication research include the completely repeated measures or "within-groups" factorial ANOVA, where each participant contributes data to each cell in the design, or a "mixed design" factorial ANOVA, where one factor is between groups while the other factor is "within-groups." We focus entirely on the between groups analysis of variance in this chapter. Entire books have been written about the analysis of data resulting from between, within, and mixed designs, and the many complicated issues that the analysis of complicated designs introduce. I refer you to one or more of the classic books on the topic, such as Keppel (1991), Keppel & Zedeck (1989), and the massive Winer, Brown, & Michels (1991) for detail on the analysis of more complicated designs.

16.2.1 Partitioning Variance in Y in a Balanced Factorial Design

Before showing how the hypothesis of interaction is tested in analysis of variance, it is worth going the process we went through in Chapter 14 of partitioning the variance of the dependent variable Y (risk judgment) into its components. The data for this exercise are presented in Table 16.1. For the purpose of illustration, I have made these data up, but they are consistent with the results reported in Berger (2000). In this hypothetical data set, 16 participants (8 men and 8 women) were randomly assigned in equal numbers to either the *Information Present* or *Information Absent* condition. The data show, for example, that the 4 men randomly assigned to the information present condition reported risk judgments of 30, 40, 20 and 30. This gives a mean for this cell of the design of $\overline{Y}_{MP} = 30$. This table also provides the *marginal means* for each factor, representing the average of the cases in that row or column of the table. So, for example, the mean risk judgment of the 8 males in the study was $\overline{Y}_M = 37.50$ and the mean risk judgment for the 8 participants who received no population trend information was $\overline{Y}_A = 50$. Finally, the table also shows that the mean risk judgment for all 16 people in the study was $\overline{Y} = 48.75$. The mean of all n units in the data is typically called the *grand mean* in the lingo of analysis of variance.

Observe in Table 16.1 that there is considerable variation in people's risk judgments around the grand mean. Some people perceived themselves to be more vulnerable than the grand mean, whereas others perceived themselves to be less vulnerable than the grand mean. Of course this isn't surprising because people will differ in how vulnerable they perceive themselves to be to crime in a community for any number of reasons. According to the logic and mathematics of between-groups factorial analysis of variance, such individual differences can be broken into several components. Much like in single-factor ANOVA, one component is how people in the different groups differ from grand mean. But here we have two ways of categorizing people into groups because we have two factors. Consider a man assigned to the population information absent condition. One of these men, call him John, had a perceived risk judgment of 40. His 40 can be attributed in part to how men differ from the grand mean $(\overline{Y}_M - \overline{Y})$ and also to how people who received no population trend information differ from the grand mean on average $(\overline{Y}_A - \overline{Y})$. But there is another source of variation that is attributable to being *both* a man *and* being assigned to the population trend information absent condition. On average, such people also differ from the grand mean in such a way that cannot be attributed merely to the additive effects of the two factors. Intuitively, we might want to symbolize this as $\overline{Y}_{MA} - \overline{Y}$, but doing so would be problematic because the size of $\overline{Y}_{MA}$ depends in part on both $(\overline{Y}_M - \overline{Y})$ and $(\overline{Y}_A - \overline{Y})$. The joint effect of being a man *and* being assigned to the information absent condition can be quantified as $\overline{Y}_{MA} - \overline{Y}_M - \overline{Y}_A + \overline{Y}$. Everything left over is individual differences between people in the same cell in the design (i.e., how John's Y differs from the other men in his condition in the study: $Y - \overline{Y}_{MA}$).

Before showing how John's score of 40 can be partitioned into these 4 components, I need to make the important distinction between a *balanced* and an *unbalanced* factorial design. In a balanced factorial design, the number of cases in each cell of the design is the same. The design in Table 16.1 is balanced because each cell contains 4 cases. In contrast, in an unbalanced design, the number of cases differs across the cells. The following discussion on apportioning variation in Y applies only to balanced designs. For unbalanced designs, the mathematics I am about to describe do not work, as I illustrate later.

Table 16.1
Hypothetical Data from Berger (2000), Balanced Design

	Population Information				
Gender	Present		Absent		Marginal Means
Male	30	40	40	50	
	20	30	50	40	$\overline{Y}_M = 37.5$
	$\overline{Y}_{MP} = 30$		$\overline{Y}_{MA} = 45$		
Female	60	60	50	60	
	80	60	60	50	$\overline{Y}_F = 60$
	$\overline{Y}_{FP} = 65$		$\overline{Y}_{FA} = 55$		
Marginal Means	$\overline{Y}_P = 47.5$		$\overline{Y}_A = 50$		$\overline{Y} = 48.75$

In a balanced design each case's Y score, expressed as deviation from the grand mean $(Y - \overline{Y})$, can be expressed as a sum of the 4 components just described. For example, for men (M) in the absent (A) condition, the following equation holds:

$$(Y - \overline{Y}) = (\overline{Y}_M - \overline{Y}) + (\overline{Y}_A - \overline{Y}) + (\overline{Y}_{MA} - \overline{Y}_M - \overline{Y}_A + \overline{Y}) + (Y - \overline{Y}_{MA})$$

So for John,

$$\begin{array}{ccl} (40 - 48.75) & = & (37.50 - 48.75) + (50 - 48.75) + (45 - 37.50 - 50 + 48.75) + (40 - 45) \\ -8.75 & = & \text{-11.25} \quad + \quad 1.25 \quad + \quad 6.25 \quad + \quad -5.00 \\ -8.75 & = & -8.75 \end{array}$$

It works. In a balanced design, this is true for every case in the data.

Because we are ultimately interested in partitioning variability across the entire data set rather than for each person, we need to quantify these sources of variation across the entire data set. This is accomplished by computing each component for each case in the data set, squaring each component, and adding each squared component across all cases, just as we did in Chapter 14 for the single-factor ANOVA. The result is a sum of squares for each component. In a balanced design with two factors A and B, the following equation holds:

$$\boxed{\begin{aligned} \sum(\overline{Y}_{ijk} - \overline{Y})^2 \quad = {} & \sum(\overline{Y}_{A_i} - \overline{Y})^2 + \sum(\overline{Y}_{B_j} - \overline{Y})^2 + \sum(\overline{Y}_{A_iB_j} - \overline{Y}_{A_i} - \\ & \overline{Y}_{B_j} + \overline{Y})^2 + \sum(Y_{ijk} - \overline{Y}_{A_iB_j})^2 \end{aligned}} \tag{16.1}$$

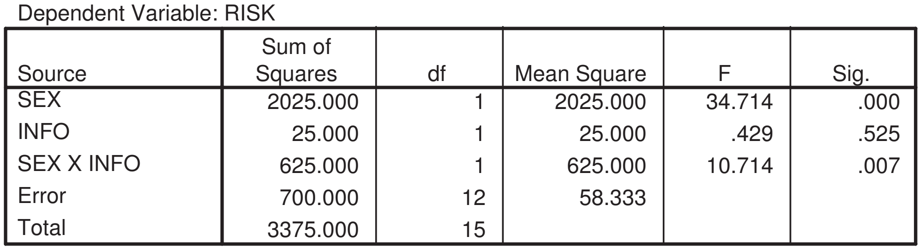
where Y_{ijk} corresponds to case k's Y measurement, with case k belonging to level i of Factor A and level j of Factor B in the analysis. The summation is over all all cases in the data file. Equation 16.1 can be rewritten symbolically as

$$\boxed{SS_{total} = SS_A + SS_B + SS_{A \times B} + SS_{error}} \tag{16.2}$$

where SS_A and SS_B are the sum of squares for the effect of factor A and factor B on Y, and $SS_{A \times B}$ is the sum of squares for the *interaction* between A and B. SS_{error} is sometimes called the *within-group sum of squares* and denoted SS_{within}. I use SS_{within} and SS_{error} interchangeably. They mean the same thing.

Figure 16.2 contains an SPSS ANOVA summary table from a factorial analysis of variance showing all these sums of squares. Observe that indeed equation 16.2 works: $3375 = 2025 + 25 + 625 + 700$. This is true because the design is balanced. It is easy to show that in a balanced design, the sources of variation described above are *independent*, in that they carry unique information about variability in Y around $\overline{Y}$. Thus, their effects can be added up as above to produce the total variation in Y, quantified as SS_{total}. But in an unbalanced design, these components are partially redundant. They carry overlapping information, and the sources of variation cannot be added up to produce total variation in Y. An example of this will be provided in section 16.2.4.

Each sum of squares also has associated with it a *Mean Square* (MS), which is computed by dividing the sum of squares by its correspondent degrees of freedom. In a factorial design, df_A is the number of levels of the A factor minus 1, df_B is the number of levels of the B factor minus 1, $df_{A \times B} = df_A \times df_B$, and $df_{error} = n - df_A - df_B - df_{A \times B} - 1$, where n is the total sample size. An ANOVA summary table such as in Figure 16.2 will also contain the degrees of freedom and MS for each source of variation.



Dependent Variable: RISK

Source	Sum of Squares	df	Mean Square	F	Sig.
SEX	2025.000	1	2025.000	34.714	.000
INFO	25.000	1	25.000	.429	.525
SEX X INFO	625.000	1	625.000	10.714	.007
Error	700.000	12	58.333		
Total	3375.000	15			

Figure 16.2 SPSS ANOVA summary table from a 2 × 2 ANOVA of the data in Table 16.1.

16.2.2 Main and Interaction Effects

Notice in Figure 16.2 that unlike in a single factor ANOVA, in factorial ANOVA, there are several F ratios, one for each of the three main components described above (the fourth component is the error component, but it has no F ratio). These F statistics are all computed by dividing the mean square for the component by MS_{error}. Each of these F ratios can be used to test different null hypotheses by computing the p-value for F.

Main Effects. In a two-factor ANOVA, there are two *main effects.* A main effect refers to the effect of one of the factors, ignoring the existence of the other factor. These main effects correspond to differences in the marginal means on the outcome variable for each factor. In this study, the two main effects are the sex main effect (males vs. females) and the population trend information main effect (present vs. absent). The sex main effect refers to the difference between the mean risk judgments of men compared to women, corresponding in these data to the difference between $\overline{Y}_M = 37.50$ (the mean risk judgment for men) and $\overline{Y}_F = 60$ (the mean risk judgment for women). The F ratio for this main effect in these data is

$$F_{Sex} = \frac{MS_{Sex}}{MS_{error}} = \frac{2025.000}{58.333} = 34.714$$

This F ratio can be used to test the null hypothesis, $H_0 : \mu_M = \mu_F$ against the alternative: $H_a : \mu_M \neq \mu_F$. The degrees of freedom for this F ratio are $df_{numerator} = df_{sex}$ and $df_{denominator} = df_{error}$, and the p-value derived from a table of critical values of F or with a computer. In these data, $F_{Sex}(1, 12) = 34.712, p < .0005$. So we can reject this null hypothesis. The obtained difference between the risk judgments of men and women is too large to attribute it to chance. It seems that women perceive themselves as more vulnerable to burglary than do men.[1]

The second main effect corresponds to the effect of population trend information on risk judgments. In these data, this main effect corresponds to the difference between $\overline{Y}_P = 47.5$ (mean risk judgment for the 8 participants who received population

[1] Of course, we really have no basis for making a statistical statement about men and women from this design, given that the participants in Berger's study were conveniently available and not obtained through any kind of random sampling plan. But with a significant F-ratio, we can discount the null hypothesis of a random process pairing respondents of different sexes to particular risk judgments, as discussed in Chapter 10.

trend information) and $\overline{Y}_A = 50$ (mean risk judgment for the eight who received no population trend information). The F ratio is

$$F_{Info} = \frac{MS_{Info}}{MS_{error}} = \frac{25.000}{58.333} = 0.429$$

and is used to test the null hypothesis $H_0 : \mu_{Absent} = \mu_{Present}$ against the alternative $H_a : \mu_{Absent} \neq \mu_{Present}$. In these data, $F(1, 12) = 0.429, p = 0.525$, so the null hypothesis cannot be rejected. It seems that giving population trend information had no effect on the participant's risk judgments. The obtained difference can most parsimoniously be attributed to "chance."

As you will see, these main effects can be misleading, depending on whether the two factors interact. Because a sensible substantive interpretation of a main effect depends on whether or not there is an interaction between the two factors, the best strategy is to first focus on the *interaction* rather than the main effects.

Interaction. If two factors A and B interact, then the effect of factor A differs across levels of factor B, and the effect of B differs across levels of factor A. In this example, if the effect of population trend information depends on whether the reader is male or female, then we would say that sex and population information interact in explaining variation in risk judgments.

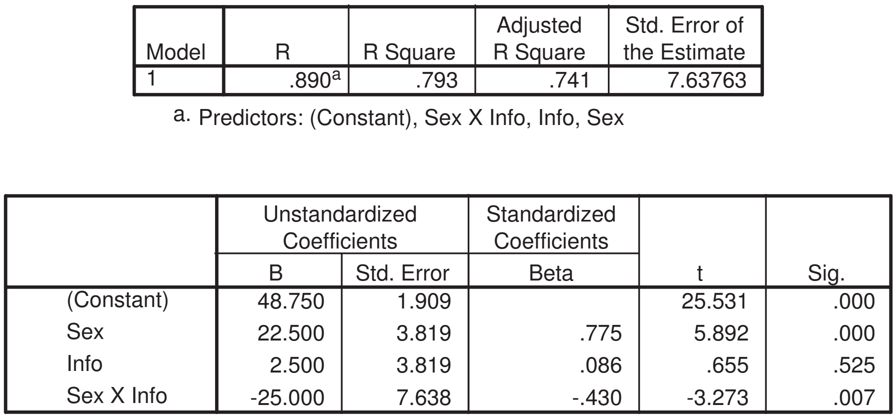
Interaction is most easily understood by considering the notion of a *simple effect.* A simple effect is the effect of one factor conditioned on the level of a second factor. For example, in this study there are 2 simple effects for population trend information. One is the simple effect of population information in men. The other is the simple effect of population information in women. In these data, the simple effect of population trend information in men is $(\overline{Y}_{MA} - \overline{Y}_{MP}) = 45 - 30 = 15$. Descriptively at least, men who had the population trend information perceived themselves less vulnerable to burglary than did men who were not given this information. The simple effect of population trend information in women is $(\overline{Y}_{FA} - \overline{Y}_{FP}) = 55 - 65 = -10$. On the surface, it would seem that giving females population trend information did not lower their perceived vulnerability. If anything, it increased it.

If 2 variables interact, then by definition of interaction, the simple effects are different. Here, we see that the simple effects descriptively are different (15 vs. -10). But we want to know whether they are *statistically* different. In other words, do we have reason to reject "chance" as the best explanation for the obtained difference between the simple effects? That is, can we reject the null hypothesis that $H_0 : (\mu_{MA} - \mu_{MP}) = (\mu_{FA} - \mu_{FP})$ in favor of the alternative $H_a : (\mu_{MA} - \mu_{MP}) \neq (\mu_{FA} - \mu_{FP})$? The F ratio gives us the key. If there is no interaction between population trend information, then we expect F to be about 1. Using information from the ANOVA summary table in Figure 16.2

$$F_{Sex \times Info} = \frac{MS_{Sex \times Info}}{MS_{error}} = \frac{625.000}{58.333} = 10.714$$

with a p-value of .007. We can reject the null hypothesis because the p-value is smaller than 0.05, $F(1, 12) = 10.714, p = .007$. It seems that population trend information and gender interact—the size of the effect of population information depends on whether the person is a male or female.

There is another way of interpreting this interaction because there are two more simple effects that we could compare. We could ask whether the differences in risk judgments between men and women differ depending on whether or not the person received population trend information. The simple effect of sex when no population



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.890[a]	.793	.741	7.63763

a. Predictors: (Constant), Sex X Info, Info, Sex

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	48.750	1.909		25.531	.000
Sex	22.500	3.819	.775	5.892	.000
Info	2.500	3.819	.086	.655	.525
Sex X Info	-25.000	7.638	-.430	-3.273	.007

Figure 16.3 Regression output corresponding to a 2 × 2 ANOVA of the data in Table 16.1.

trend information is provided is ($\overline{Y}_{FA} - \overline{Y}_{MA} = 55 - 45 = 10$). The simple effect of gender when population information is provided is $(\overline{Y}_{FP} - \overline{Y}_{MP}) = 65 - 30 = 35$. We can ask whether 35 is larger than 10 to a statistically significant degree, but we don't need to conduct another test because the F ratio for the interaction can also be used to test the null hypothesis that the simple effects are actually the same and differ from each other in the data available by just a chance mechanism. We can reject this null hypothesis and claim that females perceive themselves to be more vulnerable than men after reading the story, but more so when population trend information is provided compared to when it is absent.

When two factors interact, the main effects may not have a substantively useful interpretation, because the main effect of a factor is defined mathematically as the average simple effect of that factor. For example, notice the main effect of population trend information, $\overline{Y}_A - \overline{Y}_P = 2.50$, which is equal to the average of the two simple effects of population trend information: $(\overline{Y}_{FA} - \overline{Y}_{FP}) = 55 - 65 = -10$ and $(\overline{Y}_{MA} - \overline{Y}_{MP}) = 45 - 30 = 15$, so $(-10 + 15)/2 = 2.50$. If the simple effects are different from each other (which is what interaction is by definition), then the main effect of a factor will probably be a poor summary of the simple effects of that factor. So in the presence of a statistically significant interaction, it makes more sense to focus your interpretation on the interaction than on the main effects.

16.2.3 The Regression Equivalent of a 2 × 2 Factorial ANOVA

In the last two chapters, I made the point that ANOVA and ANCOVA are just special forms of multiple regression. The same is true for factorial analysis of variance. To illustrate, I provide the regression output from a multiple regression predicting perceived risk from population trend information, sex, and their interaction in Figure 16.3. Prior to running this analysis, I used a form of group coding called *effect coding*. For this coding scheme, the levels of the sex factor (*Sex*) were coded such that males = -0.5 and females = 0.5. Similarly, the population trend information factor (*Info*) was coded such that those assigned to the information present were assigned a score of -0.5

and those assigned to the information absent condition were given a score of 0.5. The following regression model was then calculated estimating risk judgment (Y):

$$\hat{Y} = a + b_1(Sex) + b_2(Info) + b_3(Sex \times Info)$$

where $Sex \times Info$ is a new variable defined as the product of Sex and $Info$ from the effect coding scheme described above.[2] In this model, b_1 quantifies the main effect of sex, b_2 quantifies the main effect of population trend information, and b_3 quantifies the interaction between sex and population trend information. In Figure 16.3 you will find the regression coefficients and tests of significance from this regression model. The best fitting regression model is

$$\hat{Y} = 48.750 + 22.500(Sex) + 2.500(Info) - 25.000(Sex \times Info)$$

Observe that the coefficient for $Info$ (b_2) of 2.50 is equal to the difference between the population information marginal means: ($\overline{Y}_A = 50.00$) − ($\overline{Y}_P = 47.50$) = 2.50. Furthermore, the p-value is the same as the p-value from the F ratio for the information main effect in the ANOVA (see Figure 16.2). This is because the tests are mathematically identical. Notice that the square of t for $Info$ from the regression analysis is equal to F for the information main effect in the ANOVA summary table (i.e., $-0.655^2 = 0.429$). Similarly, the coefficient for Sex (b_1) of 22.5 is exactly equal to the difference between the sex marginal means: ($\overline{Y}_F = 60.00$) − ($\overline{Y}_M = 37.50$) = 22.50, and the p-value for is the same as the p-value from the F ratio from the ANOVA, and the square of the t statistic is equal to F for the sex main effect from the ANOVA (i.e., $5.892^2 = 34.714$). Finally, it should come as no surprise now that the coefficient for the product of Sex and $Info$ (b_3) is equal to the difference between the simple effect of information for males and the simple effect of information for females: $(\overline{Y}_{FA} - \overline{Y}_{FP}) - (\overline{Y}_{MA} - \overline{Y}_{MP}) = (55 - 65) - (45 - 30) = -10 - (15) = -25$. Again, the the p-value for the regression coefficient for this product is the same as the p-value for the interaction in the ANOVA table, and $t^2 = F$.

16.2.4 Factorial ANOVA and Unbalanced Designs

In a perfect empirical world, our designs will be balanced. The effects we estimate will provide unique information about systematic variation in Y, and the four components of variance derived in section 16.2.1 will completely add up to the total variance in Y. But a factorial design often is not balanced. In experimental contexts, sometimes we have to throw out some of the data for whatever reason, producing certain cells that have a smaller number of participants. In nonexperimental contexts, factorial designs are almost certainly going to be unbalanced. For example, if we were to crossclassify respondents to a telephone poll into a 2 × 2 table (such as male vs. female and kids vs. no kids) and analyze how much TV people in these categories watch on average with a factorial ANOVA, it is highly unlikely that the four cells will contain the same number of people. When the design is unbalanced, we can still do factorial ANOVA and test for interaction, but there are few twists on the interpretation that need to be considered. Most computer programs can handle unbalanced designs as easily as balanced designs,

[2] Effect coding is typically described as a coding scheme using 1 and −1 to code various levels of the factors. However, effect codes can be multiplied by any constant without changing the results. But modification of the effect codes in this way changes the regression coefficients. By using 0.5 and −0.5 rather than 1 and −1, the regression coefficients exactly equal the differences between the marginal means.

Table 16.2
Hypothetical Data from Berger (2000), Unbalanced Design

	Population Information		Marginal Means	
Gender	Present	Absent	Unweighted	Weighted
Male	30 20 20 40 30 40	40 50 50 40	$\overline{Y}_M = 37.50$	$\overline{Y}_M = 36.00$
	$\overline{Y}_{MP} = 30.00$	$\overline{Y}_{MA} = 45.00$		
Female	60 60 80 60	50 60 40 50 60 50 60 70	$\overline{Y}_F = 60.00$	$\overline{Y}_F = 58.33$
	$\overline{Y}_{FP} = 65.00$	$\overline{Y}_{FA} = 55.00$		
Unweighted Marginal Means	$\overline{Y}_P = 47.50$	$\overline{Y}_A = 50.00$	$\overline{Y} = 48.75$	
Weighted Marginal Means	$\overline{Y}_P = 44.00$	$\overline{Y}_A = 51.67$		$\overline{Y} = 48.18$

so the real effort for you as the data analyst is to make sure you understand what the computer is telling you.

The data in Table 16.2 is largely a replication of Table 16.1, but I've added a few cases to some of the cells to produce an unbalanced design. Although the design is unbalanced, notice that the cell means are identical to the cell means in the balanced design, so I haven't done anything to the differences between the cell means. I've also added some new rows and columns that will be important in the forthcoming discussion. In Figure 16.4, panel A, I have provided the ANOVA summary table from a factorial ANOVA on these data, and in panel B you will find the output from a multiple regression equivalent using the coding scheme described in the previous section.

Partioning Variation in Y in an Unbalanced Design. Equation 16.2 says that the total variance in Y (quantified as SS_{total}) is equal to the sum of the sums of squares for the four components in this two-factor factorial ANOVA. But this applies only to balanced designs. Notice in 16.4, panel A, that the four sum of squares do not add up to the total sum of squares: $2557.895+31.579+789.474+1400.000 = 4778.948 \neq 4927.273$. This is typical when a design is unbalanced and results from the fact that in an unbalanced design, the factors and their interaction are intercorrelated variables. In this example, sex and population trend information factors carry redundant information as to the estimates of their effects. For example, the sex effect contains a part of the population information effect, because females are more likely than males to be in the population trend information absent condition in this unbalanced design.

In an unbalanced design, the sources of variation in Y cannot be derived using the logic and method described in section 16.2.1. Instead, the sums of squares can be derived much like they were derived in analysis of covariance by considering the analysis of variance as a linear regression using effect coding of the factors as was done in section 16.2.3. The sum of squares for a factor is assessed by quantifying how much the regression sum of squares decreases when a factor is excluded from the regression. Imagine running a regression estimating risk judgments Y from Sex, $Info$, and their product using coding scheme discussed in section 16.2.3. The regression sum of squares from this regression is listed in Figure 16.4 panel B.

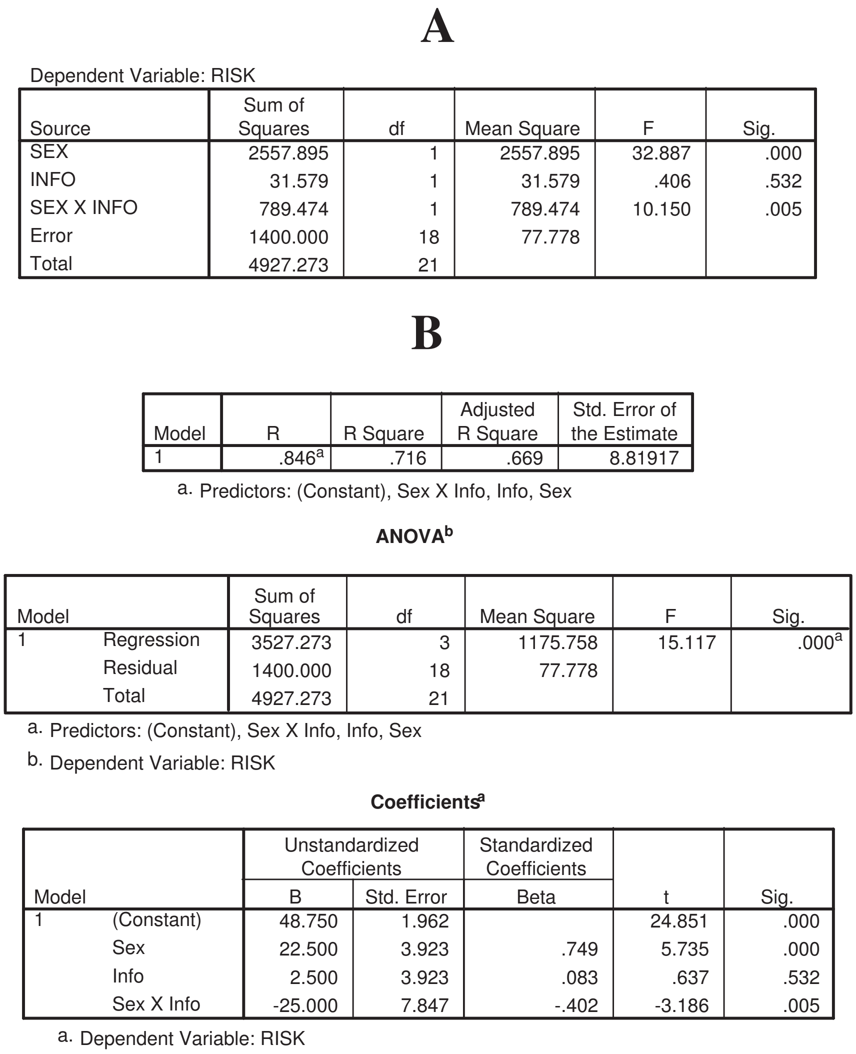
To calculate the sum of squares for a specific effect, derive the difference between sum of squares for the model that includes the main effects and the interaction and the sum of squares from a regression model that excludes just that effect. Table 16.3 contains the regression sum of squares from the regressions necessary to derive the sums of squares for each effect. For example, to calculate the sum of squares for the sex main effect, SS_{Sex}, subtract the sum of squares for the model that excludes the

Table 16.3
Regression Sums of Squares for Different Models

Factors in Model	$\text{SS}_{regression}$
Sex, Info, Sex × Info	3527.273
Sex, Info	2737.799
Sex, Sex × Info	3495.694
Info, Sex × Info	969.378

main effect for Sex (969.378 in Table 16.3) from the full model including all main effects and the interaction (3527.273). The difference is $3527.273 - 969.378 = 2557.895$, which is exactly what the output in Figure 16.4, panel A says. The method yields the following sums of squares for each effect:

$$\begin{aligned} SS_{Sex} &= SS_{Sex,Info,Sex\times Info} - SS_{Info,Sex\times Info} = 3527.273 - 969.378 = 2557.895 \\ SS_{Info} &= SS_{Sex,Info,Sex\times Info} - SS_{Sex,Sex\times Info} = 3527.273 - 3495.694 = 31.579 \\ SS_{Sex\times Info} &= SS_{Sex,Info,Sex\times Info} - SS_{Sex,Info} = 3527.273 - 2737.799 = 789.474 \end{aligned}$$



A

Dependent Variable: RISK

Source	Sum of Squares	df	Mean Square	F	Sig.
SEX	2557.895	1	2557.895	32.887	.000
INFO	31.579	1	31.579	.406	.532
SEX X INFO	789.474	1	789.474	10.150	.005
Error	1400.000	18	77.778		
Total	4927.273	21			

B

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.846[a]	.716	.669	8.81917

a. Predictors: (Constant), Sex X Info, Info, Sex

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3527.273	3	1175.758	15.117	.000[a]
	Residual	1400.000	18	77.778		
	Total	4927.273	21			

a. Predictors: (Constant), Sex X Info, Info, Sex

b. Dependent Variable: RISK

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	48.750	1.962		24.851	.000
	Sex	22.500	3.923	.749	5.735	.000
	Info	2.500	3.923	.083	.637	.532
	Sex X Info	-25.000	7.847	-.402	-3.186	.005

a. Dependent Variable: RISK

Figure 16.4 ANOVA summary table (A) and 2 × 2 ANOVA as a regression analysis (B) using the data in Table 16.2.

The remaining sum of squares, SS_{error}, is calculated as SS_{total} minus the sum of squares for the regression that includes all the main effects and the interaction. In these data, $SS_{error} = 4927.273 - 3527.273 = 1400.000$. Verify using Figure 16.4 that these sums of squares are all correct.

In an unbalanced factorial design with two factors A and B, the following equation holds:

$$\boxed{SS_{total} = SS_A + SS_B + SS_{A\times B} + SS_{error} + SS_{redundant}} \tag{16.3}$$

where $SS_{redundant}$ is variance in Y that cannot be uniquely attributed to any of the factors or their interaction because of the intercorrelation between the factors. A little algebraic manipulation tells us that

$$\boxed{SS_{redundant} = SS_{total} - (SS_A + SS_B + SS_{A\times B} + SS_{error})} \tag{16.4}$$

which in this example works out to

$$SS_{redundant} = 4927.273 - (2557.895 + 31.579 + 789.474 + 1400.000) = 148.325$$

Because $SS_{redundant}$ will be greater than 0 in an unbalanced design, it follows that in an unbalanced design, $SS_{Total} > (SS_A + SS_B + SS_{A\times B} + SS_{error})$.

The correlation between the factors that is produced by an unbalanced design does not produce a major mathematical problem (although it used to before computers were around to help out) because it is fairly easy to compute each variable's unique effect, defined as its effect after controlling for the effects of the other variables it is correlated with. Once each variable's unique sum of squares is derived as above, then the procedures for generating the mean squares, F ratios, and p-values described in section 16.2.2 can be used. But much like in multiple regression, if the factors are highly correlated, then even though one or both may have strong effects taken separately, they may have very small unique effects after the effects of the other factor and the interaction are considered and partialed out. So a small effect for one factor in an unbalanced design doesn't necessarily mean that the factor has a small effect when considered in isolation. It may simply have a small *unique* effect. In hypothesis testing contexts, what this means is that a main effect may be nonsignificant even if the variable has a large effect on the outcome measure because part of its effect is thrown out. Shared variation in explaining variation in Y is given to no factor or the interaction and instead is captured mathematically as $SS_{redundant}$.

Weighted versus Unweighted Marginal Means. To understand another important difference between balanced and unbalanced designs, you need to be understand the distinction between *unweighted* and *weighted* marginal means. Consider the mean risk judgment for males. There are 10 males in the data in Table 16.2. If you add up the 10 risk judgments provided by the males in the study, you will get 360. Divide 360 by 10 and you will get 36. This is the *weighted* mean risk judgment for males. It is a weighted because it is equivalent to the mean of the 2 male means in the table (males present and males absent) with each mean weighted by its sample size. That is, $36 = [6(30) + 4(45)]/10$. Similarly, the *weighted* mean risk judgment for females is the sum of the 12 female risk judgments divided by 12, which is $700/12 = 58.33$, which is the same as $[4(65) + 8(55)]/12$. In contrast, the *unweighted* marginal means (sometimes called *estimated means*, *e-means*, or *least squares means*), are derived by simple averaging of the means in that row or column in the table. So the unweighted mean male risk judgment is $(30+45)/2 = 37.50$, and the unweighted mean female risk judgment is $(65+55)/2 = 60$. So the difference between the marginal means depends on whether you define those means as weighted or unweighted. The difference between the weighted marginal means is 22.33, whereas the difference between the unweighted marginal means is 22.50.

This distinction between weighted and unweighted means was not made in section 16.2.2 because in a balanced design the weighted and unweighted means are the same. But when the design is unbalanced, the weighted and unweighted means are typically different and can be substantially different. In factorial ANOVA, the main effects are defined as the difference between the unweighted marginal means *not* the weighted means. So when interpreting the results of an unbalanced factorial ANOVA, you should base your interpretation on the unweighted marginal means, not the weighted means. A failure to recognize this can produce some seemingly bizarre situations that will be difficult to make sense of otherwise. For example, it is possible to get a statistically significant main effect for a variable even if the weighted marginal means are exactly the same. This can happen when there are large differences in the samples sizes across cells in the table.

As an informal proof that the main effects are tests of the difference between the unweighted marginal means and not the weighted marginal means, compare the p-values for the main effects in the ANOVA table (Figure 16.4, panel A) and the p-values for the regression weights in the regression output (Figure 16.4, panel B). Observe that they are the same, and that the F statistics in the ANOVA table are indeed the square of the t statistics from the regression output, as described earlier. Thus, these two tables show the results of mathematically identical hypothesis tests. But notice that the regression weight for sex is the difference between the unweighted marginal means, not the weighted marginal means. Similarly, the coefficient for population trend information is the difference between the unweighted marginal means, not the weighted marginal means.

The important message here is that when interpreting the results of ANOVA, it is easy (and common) to misinterpret these main effects as if they are comparisons of the means for different levels of one variable computed as if the second variable did not exist. If you compute the marginal means pretending as if the second variable did not exist in your design, you are computing the weighted means. The significance test for the main effect is not testing the difference between those weighted means. Instead, it is a test of the difference between the unweighted means. So remember, interpretations of the main effects in an unbalanced should be based on the unweighted means, not the weighted means.

16.2.5 Probing an Interaction in a Factorial ANOVA

So it seems that information about changes in the size of the population over time has a different effect on men's perceptions of vulnerability to burglary compared to women's following the reading a story about increasing crime in the community. But does that support Berger's hypothesis? Read it carefully before deciding:

> H4: Men exposed to a story showing increasing population frequencies before receiving a story depicting increasing threat during the same time period will show lower levels of victimization risk than will men receiving only a message depicting increasing threat. By contrast, among women, expo-

> sure to population increase data will not lower victimization risk levels. (Berger, 2000, p. 31–32)

Rejecting the null hypothesis of no interaction does not mean that the pattern of differences is as predicted. To assess whether the patterns of means is as predicted, it is necessary to probe this interaction. Just how this is best accomplished is a bit controversial, and there are several ways of going about it. For the sake of illustration let's focus on the balanced design in Table 16.1. The simplest approach is to analyze the simple effects separately to see if the pattern of means is consistent with the predictions. For example, a simple t test comparing perceived risk judgments in men as a function of whether or not population trend information was provided reveals that risk judgments were in fact lower on average when population trend information was provided, Welch $t(5) = 3.000, p < .05$. In women, however, population trend information seemed to have no effect on average perceived risk, Welch $t(5) = -1.732, p = .146$. Thus, just as Berger predicted, providing population trend information lowered men's perceived vulnerability to crime but not women's. This strategy of simple t tests on the simple effects uses only information provided by participants in the study that contribute to the simple effect.

There is a more powerful but somewhat more complicated strategy that you could employ. This alternative approach is to construct a focused contrast corresponding to these t tests. Using the same method as described in Chapter 14, the following contrast would quantify the difference between population trend information and no population trend information among men (from equation 14.17):

$$\delta = 1(\overline{Y}_{MA}) - 1(\overline{Y}_{MP}) + 0(\overline{Y}_{FA}) + 0(\overline{Y}_{FP}) = 15$$

with standard error estimated as (from equation 14.18):

$$s_\delta = \sqrt{58.333\left(\frac{(1)^2}{4} + \frac{(-1)^2}{4} + \frac{(0)^2}{4} + \frac{(0)^2}{4}\right)} = 5.400$$

if you assume equality of variance in risk judgments over the four cells. With this assumption, $t(12) = 15/5.400 = 2.778, p < .05$. Using the same logic for the contrast for women produces $t(12) = -10/5.400 = -1.852, p = .09$. This contrast method will tend to be somewhat higher in power than the individual t test method when the assumption of equality of variance is met. If you don't want to make this assumption, then a t test on the simple effects of interest using the Welch-Satterthwaite approach should be used.

One can sensibly ask why the test for interaction is even necessary. Berger predicted that the population trend information should reduce perceived vulnerability among men but not among women. Why not just compare the two simple effects with a series of t tests? What information is gained by testing the interaction first? A case can be made that the interaction need not be tested at all if a set of comparisons such as these do in fact turn out as expected. But you need to recognize that what is left out from this strategy is an explicit test of whether the difference between these differences is statistically different. The interaction tests the significance of the difference between the differences. Whether or not that interaction must be statistically significant in order to make the desired claim is controversial and a matter of personal opinion. In my experience, potential critics of your research will want to see a test of the interaction, even if you personally don't feel that this test is informative or necessary.

The argument is that you need to provide some formal evidence that the differences are actually different before you can claim that the simple effects actually differ. It helps a lot to think about how the hypothesis is explicitly stated. As Berger's hypothesis 4 is stated, the hypothesis does not *explicitly* predict interaction but predicts instead that the simple effect of population trend information should be zero in women but not zero in men. Interaction is *implicitly* stated in the way the hypothesis is framed. This hypothesis can be legitimately tested with two simple effects tests without testing the interaction. But most readers would expect a test that the difference in the simple effects is statistically different, and Berger indeed reported a test of that interaction, finding it to be statistically significant.

It would be unfortunate if the the pattern of simple effects is consistent with predictions when there is no evidence of interaction. This would be unfortunate because it presents a logical paradox. How can a manipulation affect one group but not the other while, at the same time, there be no evidence that the effect differs between the groups? Unfortunately, such paradoxes arise in statistics all the time. For example, it is possible for a multiple correlation to not be statistically different from zero even if some of the partial regression weights are different from zero or for the omnibus null hypothesis to be rejected in ANOVA but to find that no means differ from each other to a statistically significant degree when all possible pairwise comparisons are conducted. The test you should focus on when interpreting an analysis is the test that is most directly relevant to the question you are trying to answer. If a hypothesis explicitly predicts an interaction, it should be tested and found to be statistically significant before you cam claim the prediction is supported in the data. But if the hypothesis doesn't explicitly state an interaction and instead proposes a pattern of simple effects, then whether or not the interaction needs to be statistically significant is controversial and a matter of personal opinion.

I have focused exclusively on the relatively simple 2 × 2 between-groups design. More complicated designs are possible. For example, one or more of the factors might have more than 2 levels. Although the conceptualization of interaction as inconsistent simple effects doesn't change, probing a significant interaction can become quite a bit more complicated. Consider for example two factors, each with 3 levels. A significant interaction can still be interpreted as simple effects that are statistically different. But each simple effect is based on 3 levels of the second factor, and there are three of these simple effects. It becomes necessary not only to probe which simple effects differ from which but also which means within a simple effect are statistically different from each other. The number of possible tests required to probe the interaction can become quite large very quickly.

Another complication involves the addition of a third or even fourth factor. When there are more than 2 factors, then it becomes possible to assess 3-way, 4-way, or even 5-way interaction. If there is a 3-way interaction between X, Z, and W, this means that the interaction between, for example, X and Z differs across levels of W. In other words, a 3-way interaction implies that two or more differences between differences are themselves different. Interpreting three-way interactions can become quite complicated, and interactions higher than the third order become nearly impossible to interpret. Nevertheless, some theories and hypotheses tested by communication researchers involve questions about three way interactions, necessitating such analyses in order to test the theory or hypothesis. I do not discuss the analysis of such designs in this book, and I refer to you more advanced books on analysis of variance such as Keppel (1991).

16.2.6 Quantifying Effect Size

In Chapters 14 and 15, several effect sizes in single factor ANOVA were introduced. When there is more than one factor, how to measure effect size becomes a bit ambiguous because there are several ways of talking about effect size, just as was the case in ANCOVA. One may ask what proportion of the *total* variance in the dependent variable Y is uniquely attributable to a specific independent variable of interest. Another conceptualization of effect size quantifies the size of the effect as the proportion of variance in the outcome remaining after partialing out the other effects on Y that is uniquely attributable to the effect of interest. You may recognize these as the distinction between the squared semipartial correlation and the squared partial correlation in Chapter 13, or η^2 and partial η^2 from Chapter 15. Focusing on η^2 and partial η^2:

$$\boxed{\eta^2 = \frac{SS_{effect}}{SS_{total}}} \tag{16.5}$$

$$\boxed{\text{partial } \eta^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}} \tag{16.6}$$

where SS_{effect} is the sum of squares for the variable for which the effect size measure is desired. Plugging the numbers in from the the unbalanced design (Figure 16.4, panel A),

$$\text{SEX}: \quad \eta^2 = \frac{2557.895}{4927.273} = 0.519; \quad \text{partial } \eta^2 = \frac{2557.895}{2557.895 + 1400.000} = 0.646$$

$$\text{INFO}: \quad \eta^2 = \frac{31.579}{4927.273} = 0.006; \quad \text{partial } \eta^2 = \frac{31.579}{31.579 + 1400.000} = 0.022$$

$$\text{SEX} \times \text{INFO}: \quad \eta^2 = \frac{789.474}{4927.273} = 0.160; \quad \text{partial } \eta^2 = \frac{789.474}{789.474 + 1400.000} = 0.361$$

Choosing between these is sometimes difficult, and which is the correct measure of a factor's effect on Y is controversial. Contrary to popular belief, both η^2 and partial η^2 can be affected by the size of the effect of the other factors on Y, as well as how large the interaction is. Partial η^2 will tend to be more affected by the size of the other effects and will generally be larger than η^2. So if you want a measure of effect size for a factor in a factorial ANOVA that is less affected by the size of the other effects in the analysis, use η^2 rather than partial η^2. But both measures of effect size are affected by the intercorrelation between effects that occurs when a design is unbalanced.

In my judgment, partial η^2 is not a good measure of effect size and should not be used. The primary problem with partial η^2 is that an investigator can make it nearly as large as desired by increasing the complexity of the research design. By increasing the number of factors in an analysis of variance that have *some* effect on Y, partial η^2 for every variable will tend to increase because it gauges the unique effect of a variable relative to variance in Y left unexplained by the other factors. By contrast, η^2 indexes a variable's unique effect relative to *total* variance in Y. As a result, η^2 is not nearly so influenced by the number of factors in the design. To be sure, η^2 can be affected by the inclusion of additional variables if they are intercorrelated with the effect of interest, but this will *lower* η^2, not increase it. In addition, η^2 is conceptually equivalent to the change in R^2 in hierarchical regression—a measure of a variable's effect that is widely used in communication. That is, η^2 can be thought of as the incremental proportion of variance in Y that is explained by including that factor in the analysis relative to when it is excluded. Partial η^2 does not have this nice interpretation.

Because both of these measures of effect size depend in part on the other variables in the analysis, how intercorrelated the variables are, and the effect of those other variables on Y, it is hard to compare effect sizes across studies that differ in design. For example, suppose investigators A and B are both interested in the effect of online versus traditional print news on public affairs knowledge and conduct similar studies at the same time. They use the same measure of public affairs knowledge (a 20–item multiple choice test of knowledge of recent world events, Y) and a measure of form of news exposure (online versus print, X), and each study is based on 100 participants. Investigator A's study is the simplest, including only the single independent variable X manipulated in an experimental design, where participants are randomly assigned to be exposed to either a print or online version of a newspaper for 30 minutes, after which they are given a test of information contained in the news. Investigator B manipulates X in exactly the same way as A but has a second independent variable W crossed with X in a factorial design. Suppose W is a number of exposures manipulation, operationalized as the number of sessions of exposure the participant receives (30 minutes over one day or 30 minutes over 3 days, 10 minutes each). In short, both studies are identical with the exception of an additional manipulation in Investigator B's study and, of course, different participants. Each investigator reports a common measure of the effect of X on Y.

Regardless of whether A and B consistently report η^2 or partial η^2, their effect sizes are not necessarily comparable. Suppose for example that both report $\eta^2 = 0.20$. Without more information, we cannot necessarily say that X has the same effect on Y in these studies even though X appears to be explaining the same proportion of variance in Y. Because investigator B manipulated a second variable W, that manipulation as well as the interaction between X and W may increase variability in Y, with the amount of that increase being a function of how large those effects are. For example, distributing the same learning time over more sessions could increase the number of relatively high learning scores in B's study compared to what A observed because some (but not necessarily all) of the participants might be less fatigued over three short learning periods. So the total variance of Y (quantified as SS_{total}) may be quite a bit higher in B's study even though they have the same sample sizes. So an η^2 of 0.20 corresponds to more variability in learning explained by X in investigator B's study, even though X explains the same amount of relative variability (i.e., $SS_{effect(X)}/SS_{total}$). Without knowing more about between study differences in the variance of Y, the effect sizes cannot be meaningfully compared. Changing to partial η^2 does not solve the problem. Indeed, partial η^2 is even less comparable across these studies because partial η^2 quantifies the proportion of the variance in Y remaining after partialing out the other variables that X uniquely explains. Because investigator A included no other variables in the analysis, $\eta^2 =$ partial η^2. Just by including an additional independent variable in the design that has some effect on Y, the effect of X on Y increases in B's study using this measure of effect size. Partial η^2 is determined in part by the number of additional variables in the analysis and so isn't comparable across studies that differ in the number or nature of the additional variables. Had A included additional variables (manipulated or just measured) related to Y, partial η^2 likely would have been larger.

But even if W and $X \times W$ had absolutely no effect on Y in B's study (meaning they did not affect either the means or the total variability in Y relative to variability observed in A's study), the meaning of η^2 might be different in the two studies. Imagine that in B's study, the sample size in the print-multiple exposure condition was smaller than in the other three cells, perhaps because participants in this condition found the study less interesting and were less likely to return for the second or third exposure period. In that case, the independent variables (and their interaction) are intercorrelated. Most discussions of effect size in the communication and other literatures have assumed that the total sum of squares in an experiment can be partitioned perfectly into nonoverlapping components, as reflected in Levine and Hullet's (2002) examples and claim that "η^2 has the property that the effects for all components of variation (including error) will sum to 1.00" (p. 619). But life in science is not always so clean and perfect. Even in true experiments where the investigator has some control over the intercorrelation between variables through random assignment and control of cell sizes, things happen that induce correlation between the independent variables, such as procedural errors, discarding of participants due to suspicion about a deception, equipment malfunctions, and so forth. Unless there is some attempt to reequalize cell sizes (which introduces new design and analysis problems and can't generally be recommended), it becomes impossible to perfectly partition total variance into the effects of interest plus error. In this case, η^2 will be reduced in study B in proportion to how predictable X is from W and $X \times W$. Keep in mind that η^2 quantifies the proportion of *total* variance in Y uniquely attributable to X. When independent variables are correlated, some of the variance in Y that X might explain had X been the only factor in the analysis is not attributed to X statistically (or any other variable for that matter) because variability in Y attributable to more than one independent variable is eliminated from η^2 (and partial η^2 as well). Because A's study has only a single independent variable, this does not affect the interpretation of η^2 in that study. The fact that η^2 is the same in B's study in spite of the intercorrelation between W and X suggests that X may have a larger effect on Y in B's study, but it is impossible to know just how much larger. Using partial η^2 does not eliminate this ambiguity in the comparison of effect sizes, as it too is affected by the intercorrelation between independent variables.

16.3 Moderated Multiple Regression

In Berger's study, both of the independent variables were categorical. But sometimes a research design includes two groups (e.g., men and women, or participants in experimental or control group) both measured on a second *quantitative* variable, such as a personality variable or some other quantitative dimension. The researcher may be interested in whether the quantitative variable is related to the outcome variable differently in the two groups or whether the average difference between the groups on the outcome variable depends on the values of the quantitative variable. Or the researcher might have two quantitative variables and is interested in knowing if the relationship between one of those variables and the outcome varies systematically as a function of the values of the second variable.

Communication researchers sometimes approach the analysis of data from a design of this sort by categorizing one or both of the quantitative independent variables in some fashion and then subjecting the dependent variable to a factorial analysis of variance. For example, the researcher might place each case into a "high" or a "low" group based on whether the case's score is above or below the sample mean or median on one or both of the quantitative independent variables. But it is not necessary to categorize in this fashion, and doing so is an inefficient way to use the data available that I strongly discourage for reasons discussed in Section 16.5 (also see Bissonnette, Ickes, Berstein, & Knowles, 1990; Irwin & McClelland, 2001; Irwin & McClelland, 2003; MacCallum, Zhang, Preacher, & Rucker, 2002; Streiner, 2002; Veiel, 1988). A much more efficient approach is *moderated multiple regression*. In a moderated multiple regression, the goal is to assess whether the regression coefficient for a predictor variable in a model varies as a function of the values of a second predictor variable. It is worthwhile to compartmentalize this discussion as a function of the levels of measurement of the variables presumed to be interacting (i.e., all quantitative or one categorical and the other quantitative), although as you will see many of the same interpretational principles apply regardless.

16.3.1 Interaction Between a Dichotomous and a Quantitative Variable

Consider a simple hypothetical study similar conceptually to Monahan and Lannuitti (2000). The data from this hypothetical study will be used throughout this section, and the details about the data file used to generate the analyses here can be found in Appendix E9 on the CD. The question motivating this study is whether alcohol use moderates the relationship between a man's self-esteem and his willingness to engage in self-disclosure. An individual difference such as self-esteem may be less related to self-disclosure after drinking because alcohol serves as a "social lubricant," easing social anxiety during conversations and thereby reducing the effect of an individual difference such as self-esteem on social interaction. To conduct this study, the researcher recruited 40 men individually to a laboratory. Upon arrival at the laboratory, each man was given a self-report measure of self-esteem, with possible scores ranging between 1 and 5 (variable SE). Each participant was then placed by himself in a room for 60 minutes containing a keg of beer, a two-liter bottle of soda, a couch, a newspaper and several magazines, and a television. The participants randomly assigned to the *alcohol condition* ($C = 1$ in the data) were told that the investigator was interested in how people respond to new social encounters and how alcohol may affect those responses. During the 60 minute period, the participants were told they could watch television, relax or read, and that they were free to drink as much beer as desired from the keg. Participants randomly assigned to the *control condition* ($C = 0$ in the data) were treated identically, except they were told that the keg of beer was for a staff party later that day and not to drink anything from the keg. After the 60 minute period, the participants were escorted to another room that contained a female confederate of the experimenter. The experimenter gave them a task that they were to accomplish together (putting together a 50-piece jigsaw puzzle), and the female was instructed to flirt with the male during this task. These interactions were videotaped. Two coders then coded how much the man self-disclosed to the female, with self-disclosure scores (Y) ranging from 0 to 9.

The data for this example were constructed using a formula that produced a different relationship between self-esteem and self-disclosure in the two conditions. If high and low self-esteem groups are created using a median split and the data then subjected to a 2 × 2 ANOVA using the procedures described in section 16.2, only a main effect of alcohol use and a main effect of self-esteem is found. The interaction is not significant, $F(1, 36) = 2.046, p = .16$. This analysis would lead to the conclusion that the effect of self-esteem on self-disclosure does not vary depending on a person's alcohol use. Alternatively, this same lack of interaction means that alcohol use has the same effect on self-disclosure regardless of a person's self-esteem. But I strongly discourage this approach to analyzing the data from this study, for reasons discussed in section 16.5.

An analysis of the same data using moderated multiple regression, the procedure described in this section, yields a different but correct the conclusion. Figure 16.5 provides a scatterplot and the least squares regression lines for the control ($C = 0$) and alcohol groups ($C = 1$). As can be seen, there is a relationship between self-esteem and self-disclosure among students that did not drink alcohol. The regression weight for self-esteem in the control group, which I will symbolize as $b_{SE|C=0}$, is $0.663, t(19) = 2.919, p = .009$. But for students who were allowed to drink, the relationship appears different. Indeed, a formal hypothesis test reveals no statistically significant relationship between self-esteem and self-disclosure in the alcohol group. The regression weight for self-esteem in this group, $b_{SE|C=1}$, is $-0.087, t(19) = -0.350, p = .731$. Descriptively at least, the relationship between self-esteem and self-disclosure depends on whether a person has been drinking. But a formal test of interaction would allow us to rule out the possibility that the obtained difference in the regression coefficients is just "chance." In moderated multiple regression, this is easily accomplished by estimating the coefficients for the regression model below:

$$\hat{Y} = a + b_C(C) + b_{SE}(SE) + b_{C \times SE}(C \times SE)$$

where $\hat{Y}$ is a case's estimated self-disclosure and $C \times SE$ is a new variable defined as the product of a participant's self-esteem (SE) and the condition he was assigned to in the study (C). The latter term in the above equation, $b_{C \times SE}$, is sometimes called the *interaction term.* The variables that constitute the interaction, in this case C and SE, are sometimes called the *lower-order* variables in the model and the coefficients the *lower-order* effects.

The results of this regression analysis are displayed in Figure 16.6. The model is:

$$\hat{Y} = 3.094 + 2.922(C) + 0.663(SE) - 0.750(C \times SE)$$

For the question as to whether alcohol moderates the effect of social self-esteem on self-disclosure, the relevant section of the output is the size of and significance test for $b_{C \times SE}$, which here is -0.750 with a p-value of .034. This regression coefficient is statistically different from zero, so we conclude that alcohol use and social self-esteem interact in explaining variation in self-disclosure. Or we can say that self-esteem moderates the effect of alcohol on self-disclosure, or that alcohol moderates the effect of self-esteem on self-disclosure.

To illustrate just why $b_{C \times SE}$ quantifies the interaction, it is helpful to break down this model and assess the meaning of each regression coefficient. Let's start with b_{SE}. Remember that in a multiple regression *without* an interaction term, the partial regression weight for predictor variable i quantifies the estimated difference in Y between two people who differ by one measurement unit on variable i but who are equal on all the other predictor variables in the model. But when an interaction term is in a regression model, this changes the interpretation of the coefficients for the lower order variables. Consider the case where $C = 1$ and $SE = 3$, and thus $C \times SE = 3$. The regression equation yields

$$\hat{Y} = 3.094 + 2.922(1) + 0.663(3) - 0.750(3) = 5.755$$

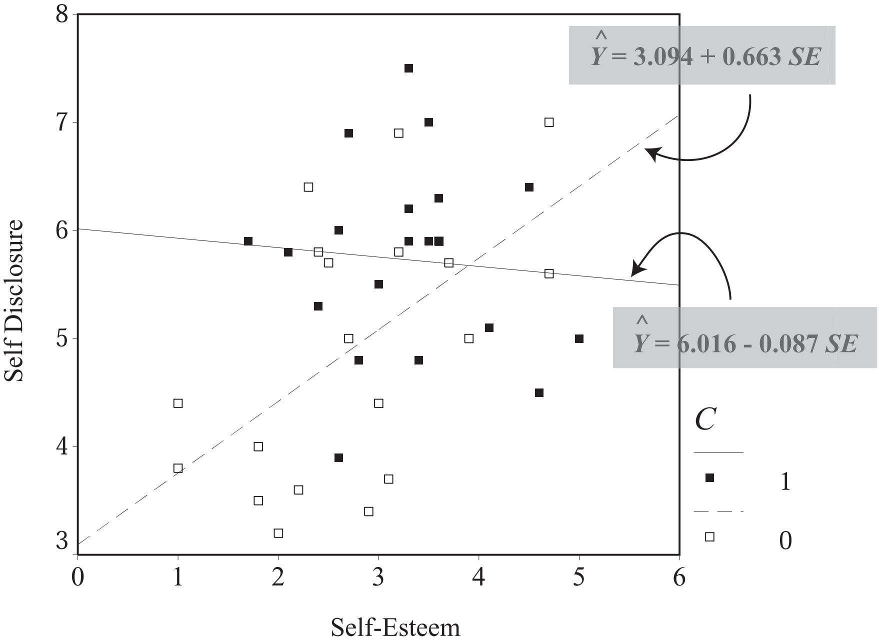


Figure 16.5 Scatterplot of the relationship between self-esteem and self-disclosure.

Keeping C constant but increasing SE to 4, the model gives

$$\hat{Y} = 3.094 + 2.922(1) + 0.663(4) - 0.750(4) = 5.668$$

Clearly, keeping C constant but increasing SE by one unit has not resulted in an increase of 0.663 in estimated self-disclosure. Instead, the estimated difference is $5.668 - 5.755 = -0.087$. Now repeat these computations, except this time using $C = 0$. In this case, when $SE = 3$, the model yields

$$\hat{Y} = 3.094 + 2.922(0) + 0.663(3) - 0.750(0) = 5.083$$

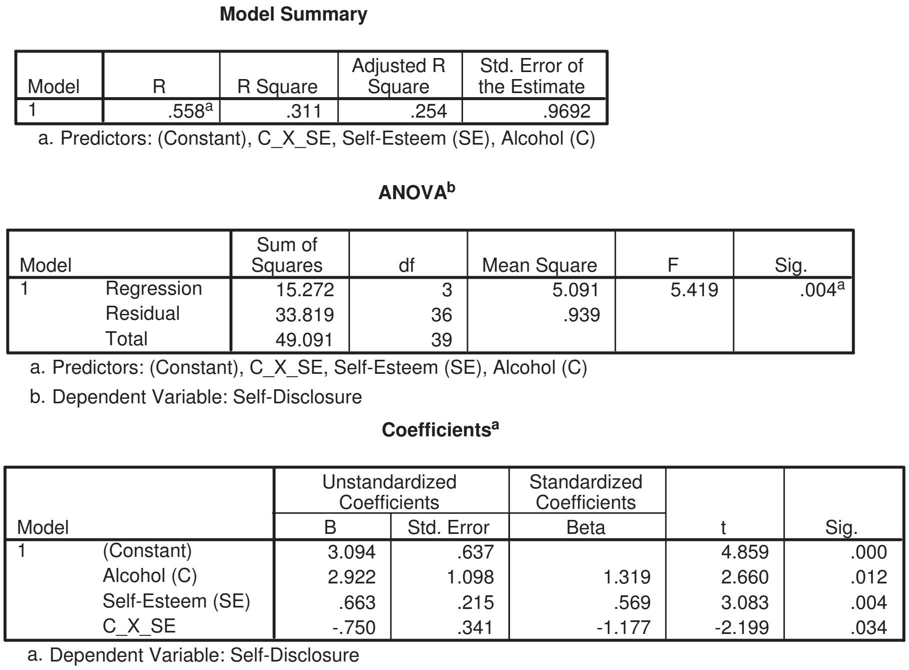
and when $SE = 4$, the model gives

$$\hat{Y} = 3.094 + 2.922(0) + 0.663(4) - 0.750(0) = 5.746$$

This time, the difference is $5.746 - 5.083 = 0.663$, which is b_{SE}. Regardless of which values of SE we choose, it would be the case that the predicted difference in self-disclosure associated with a one unit difference in SE is 0.663 when $C = 0$. So b_{SE} is the regression weight for self-esteem estimating self-disclosure when $C = 0$. In other words, it is the regression weight for SE for the control group in this study and exactly what we found when we analyzed the control group separately.

Recall that when $C = 1$, a one unit difference in SE was associated with a difference of -0.087 in estimated self-disclosure. It is no coincidence that this is exactly equal to the regression weight for the alcohol group, $b_{SE|C=1}$, from the earlier analysis. Notice as well that this difference of -0.087 is equal to $b_{SE} + b_{C \times SE}$. So the regression weight for SE when $C = 1$ is equal to $b_{SE} + b_{C \times SE}$. If $b_{SE} = b_{SE|C=0}$ and $b_{SE} + b_{C \times SE} = b_{SE|C=1}$, then simple algebra tells us that

$$b_{C \times SE} = b_{SE|C=1} - b_{SE|C=0}$$



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.558[a]	.311	.254	.9692

a. Predictors: (Constant), C_X_SE, Self-Esteem (SE), Alcohol (C)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15.272	3	5.091	5.419	.004[a]
	Residual	33.819	36	.939		
	Total	49.091	39			

a. Predictors: (Constant), C_X_SE, Self-Esteem (SE), Alcohol (C)

b. Dependent Variable: Self-Disclosure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.094	.637		4.859	.000
	Alcohol (C)	2.922	1.098	1.319	2.660	.012
	Self-Esteem (SE)	.663	.215	.569	3.083	.004
	C_X_SE	-.750	.341	-1.177	-2.199	.034

a. Dependent Variable: Self-Disclosure

Figure 16.6 SPSS output from a moderated multiple regression estimating self-disclosure from alcohol condition, self-esteem, and their interaction.

In other words, $b_{C\times SE}$ is the difference between the regression weight for self-esteem in the control group and the regression weight for self-esteem in the alcohol group. Rephrased, it can be interpreted as how the regression coefficient for self esteem changes with a one unit change in C. Indeed, observe that as C increases by one unit, the coefficient for SE changes by -0.750 (from 0.663 to -0.087). The significance test for $b_{C\times SE}$ tests the null hypothesis that this difference is attributable to "chance." Rejection of this null hypothesis means that the relationship between self-esteem and self-disclosure is statistically different in the two groups.

But what about b_C? Earlier I stated that b_{SE} is the regression weight for self-esteem estimating self-disclosure when $C = 0$. Using a similar logic, b_C is the regression weight estimating self-disclosure from alcohol use when $SE = 0$. So a one unit difference in C is associated with a 2.922 difference in estimated self-disclosure when $SE = 0$. The positive sign tells us that estimated self-disclosure is higher for people assigned to the alcohol group ($C = 1$) than the control group ($C = 0$). A visual examination of Figure 16.5 shows that when $SE = 0$, the two regression lines are indeed separated by just about 3 units. Finally, a, the regression constant, represents the estimated self-disclosure when both SE and C are equal to zero.

This example illustrates some general principles. In any regression model of the form

$$\hat{Y} = a + b_X(X) + b_W(W) + b_{XW}(XW)$$

where XW is the product of X and W, b_X represents the estimated effect of a one unit difference in X on Y when variable $W = 0$, b_W represents the estimated effect of a one unit difference in W on Y when $X = 0$, and b_{XW} represents how much b_X changes with a one unit increase in W or, conversely, how much b_W changes with a one unit increase in X. So in a regression model that includes X, W, and $X \times W$, the regression coefficients for X and W are *conditional regression weights*, *conditional effects*, or *local terms* (Darlington, 1990) and cannot be interpreted like main effects are interpreted in an analysis of variance, nor can they be interpreted as a partial regression weight would in a model without an interaction term. Instead, these regression coefficients quantify the effect of one predictor variable on the outcome variable conditioned on the other predictor variable being zero. This is important to keep in mind because it is possible that one or more of the lower order regression coefficients will have no sensible substantive interpretation whatsoever in a study if 0 is not a possible measurement on one of the predictor variables. In this example, self-esteem was measured on a scale from 1 to 5. Zero was not a possible score, so b_C and its test of significance has no meaningful interpretation here. And because SE cannot equal zero, the regression constant and its test of significance also has no substantive interpretation.

It is important to point out that the interpretational principles described here apply only to *unstandardized* regression coefficients. It is common for researchers to report standardized regression coefficients when reporting a moderated multiple regression model, but standardized regression coefficients in moderated multiple regression do not have the properties described here. Standardized regression coefficients are hard to interpret in this context, and I like to be able to apply the principles discussed above when I interpret other researchers' models, something that can't be done when standardized coefficients are reported. To ease interpretation by others, I suggest that if you feel you must report standardized regression coefficients (something I generally don't encourage), provide the unstandardized coefficients as well.

Probing a Significant Interaction. When testing for interaction with a factorial ANOVA, a significant interaction is typically followed by a simple effects analysis, where the investigator examines the effect of one independent variable at each level of the other independent variable. For example, had these data been analyzed after dichotomizing self-esteem at the median and a significant interaction found, the standard practice would be to do a simple effects analysis by either (a) examining the effect of the alcohol manipulation among people who are either "high" or "low" in self-esteem or (b) examining differences in self-disclosure as a function of self-esteem in each condition. But how would such an analysis be accomplished in moderated multiple regression given that self-esteem is not a categorical variable?

Before probing the interaction statistically, it is worth graphically representing the regression model by generating estimated values from the model using various values of the predictors. Moderated multiple regression is very holistic in its approach to assessing interactions, and much of the beauty of the method can be hidden by the mathematics. A picture can say a lot about what is happening in the data, so I strongly encourage you to first generate a set of $\hat{Y}$ values from the model and then plot them in the form of a scatterplot. In this example, this would be accomplished by first setting C to 0 and then plugging in several different values of SE into the regression formula. Repeat this process for the same values of SE but setting $C = 1$. Then generate a scatterplot, placing $\hat{Y}$ on the Y axis, SE on the X axis, and using different symbols in the plot for different values of C (see Figure 16.7).

There are two approaches to statistically probing this interaction. First, you could look at the regression weights separately in the two groups defined by the dichotomous variable. Recall that b_{SE} was interpreted as the conditional effect of self-esteem for those in the control group. So the regression output already provides information about the relationship between self-esteem and self-disclosure in the control group. We know from the output in Figure 16.6 that in the control condition, there is a positive

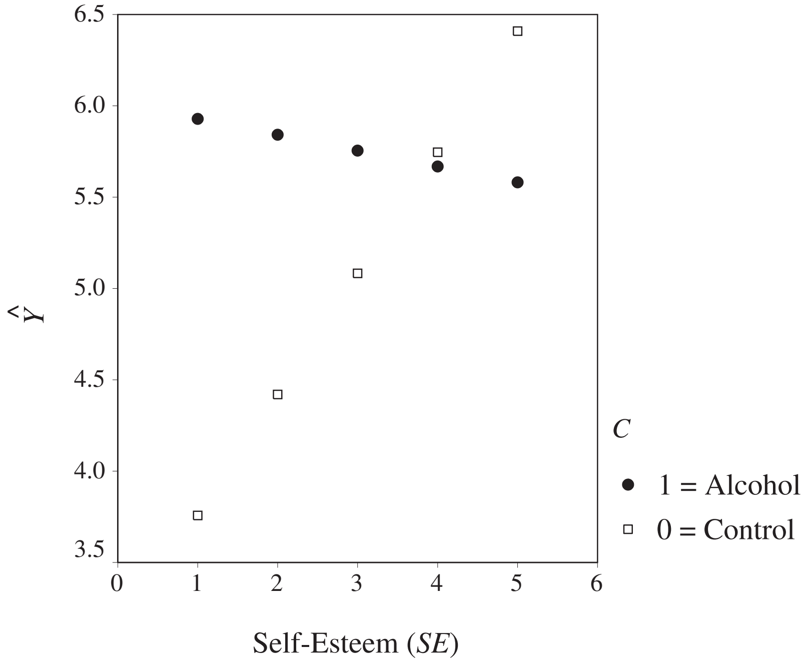


Figure 16.7 A visual representation of the interaction between self-esteem and alcohol consumption on self-disclosure.

and statistically significant relationship between self-esteem and self-disclosure, $b = 0.663, t(36) = 3.083, p = 0.004$. By rerunning the analysis, reversing the coding of C in the original data (so that $C = 0$ for the alcohol group and $C = 1$ for the control group) before computing $C \times SE$, we could get the conditional effect of self-esteem for those in the alcohol group. If you did this, you'd find $b = -0.087, t(36) = -0.350, p = .731$. So self-esteem is related to self-disclosure in the control condition but not in the alcohol condition.

But how do you assess the other kind of conditional effect—the effect of alcohol on self-disclosure at different self-esteem values? This can be accomplished using the interpretational principles outlined above. First, pick two or more "representative values" of SE at which to examine the alcohol effect and then transform the original data so that the regression output provides a test of the conditional effect of alcohol use at those values of SE. There are no hard and fast rules for selecting representative values. Aiken and West (1991) suggest using one standard deviation above the mean, the mean, and one standard deviation below the mean on one of the predictors. Or you could choose values that make some conceptual sense, or that have some kind of practical meaning. In these data, $\overline{SE} = 3.033, SD = 0.963$. One standard deviation below the mean is $3.033 - 0.963 = 2.070$ and one standard deviation above the mean is $3.033 + 0.963 = 3.996$. To test the effect of alcohol when $SE = 2.070$ (one standard deviation below the mean), create a new variable, SE', defined as $SE' = SE - 2.070$ as well as a new interaction term, $C \times SE'$, and then reestimate the regression model. The resulting regression model (see Table 16.4) is

$$\hat{Y} = 4.666 + 1.369(C) + 0.663(SE') - 0.750(C \times SE')$$

Table 16.4
Probing an Interaction in Moderated Multiple Regression

		Coeff.	s.e.	t	p
$SE' = SE - 2.070$					
	Constant	4.466	0.266	16.809	$< .001$
	Alcohol (C)	1.369	0.469	2.919	0.006
	SE'	0.663	0.215	3.083	0.004
	$C \times SE'$	-0.750	0.341	-2.199	0.034
$SE' = SE - 3.033$					
	Constant	5.104	0.223	22.870	$< .001$
	Alcohol (C)	0.647	0.318	2.036	0.049
	SE'	0.663	0.215	3.083	0.004
	$C \times SE'$	-0.750	0.341	-2.199	0.034
$SE' = SE - 3.996$					
	Constant	5.743	0.339	16.953	$< .001$
	Alcohol (C)	-0.075	0.445	-0.169	0.867
	SE'	0.663	0.215	3.083	0.004
	$C \times SE'$	-0.750	0.341	-2.199	0.034

Notice that $b_{SE'}$ and $b_{C \times SE'}$ are the same as b_{SE} and $b_{C \times SE}$, but b_C and a are different. Prior to this transformation of SE, b_C quantified the estimated difference between the control and alcohol groups when $SE = 0$. That interpretation still applies, but with this transformation, $SE' = 0$ when $SE = 2.070$, or one standard deviation below the sample mean SE in the data. So b_C can be interpreted as the effect of alcohol use for people one standard deviation below the sample mean SE. In this model, b_C is 1.369, $t(36) = 2.919, p = .006$. So when $SE = 2.070$, a one unit difference in C is associated with a statistically significant difference of 1.369 in self-disclosure, with the alcohol group having the higher expected self-disclosure score at this value of self-esteem (because the coefficient is positive).

Repeating the procedure setting $SE' = SE - 3.033$, yields (see Table 16.4)

$$\hat{Y} = 5.104 + 0.647(C) + 0.663(SE') - 0.750(C \times SE')$$

In this model, b_C quantifies the effect of alcohol when $SE' = 0$, but $SE' = 0$ when $SE = 3.033$, so the effect of alcohol at the mean SE is $b_C = 0.647, t(36) = 2.036, p = .049$. So alcohol results in greater self-disclosure at the sample mean SE. Finally, defining SE' as $SE - 3.996$, the regression model is (from Table 16.4 panel C):

$$\hat{Y} = 5.743 - 0.075(C) + 0.663(SE') - 0.750(C \times SE')$$

At one standard deviation above the sample mean SE, the effect of alcohol use is not statistically different from zero, $b_C = -0.075, t(36) = -0.169, p = .867$ (Table 16.4)

16.3.2 Interaction Between Two Quantitative Variables

In the previous section, I described the application of moderated multiple regression to testing and probing an interaction between a dichotomous and a quantitative predictor variable. In this section, the logic of moderated multiple regression is applied to the testing of interaction between two quantitative variables. As you will see, the general rules described previously apply to this situation.

Consider a slight modification to this study. In this variation, there was no experimental manipulation. Instead, all participants were simply placed in the room and told that they could drink as much beer as desired during the 60–minute period. Otherwise, the procedure was identical except that at the end of the 60–minute period, each participant's blood alcohol content was measured (variable BAC in the data set; see Appendix E9 on the CD) with a breathalyzer. Interest in this variation of the study still focuses on whether alcohol consumption moderates the relationship between self-esteem and self-disclosure, but alcohol consumption is not experimentally manipulated here. Instead, it is operationalized as the percent of a participant's blood content that is alcohol at the end of the 60-minute period.

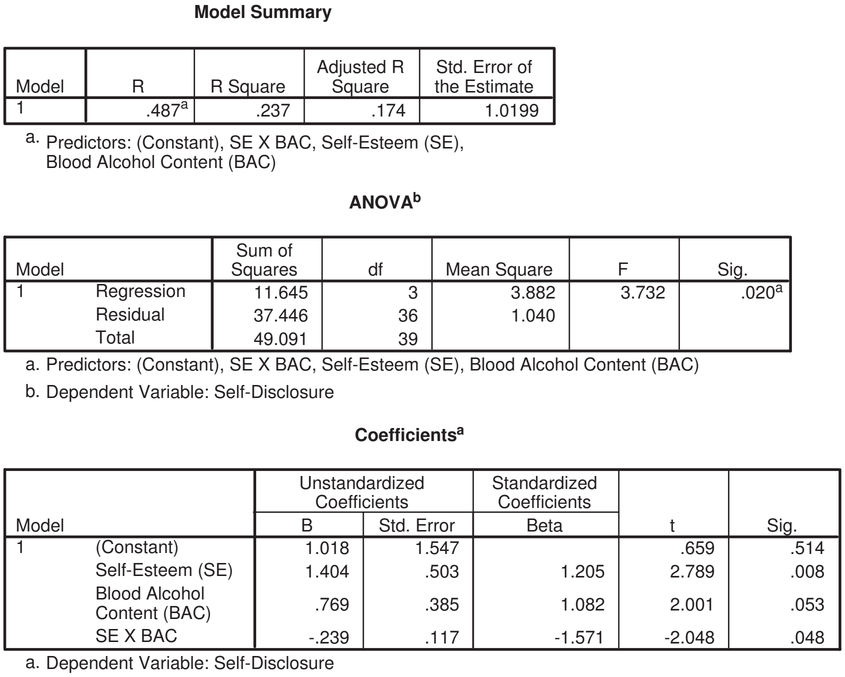
To test this question, the following model is estimated:

$$Y = a + b_{BAC}(BAC) + b_{SE}(SE) + b_{SE \times BAC}(SE \times BAC)$$

where $SE \times BAC$ is the product of blood alcohol content and self-esteem. The results of this regression are presented in Figure 16.8. The regression model is:

$$\hat{Y} = 1.018 + 0.769(BAC) + 1.404(SE) - 0.239(SE \times BAC)$$

As can be seen from the computer output, the interaction term is significantly different from zero, meaning that self-esteem and alcohol consumption interact in explaining self-disclosure. The coefficient for the interaction ($b_{SE \times BAC}$) tells us that the regression weight estimating self-disclosure from self-esteem decreases by 0.239 as blood alcohol



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.487[a]	.237	.174	1.0199

a. Predictors: (Constant), SE X BAC, Self-Esteem (SE), Blood Alcohol Content (BAC)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.645	3	3.882	3.732	.020[a]
	Residual	37.446	36	1.040		
	Total	49.091	39			

a. Predictors: (Constant), SE X BAC, Self-Esteem (SE), Blood Alcohol Content (BAC)

b. Dependent Variable: Self-Disclosure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.018	1.547		.659	.514
	Self-Esteem (SE)	1.404	.503	1.205	2.789	.008
	Blood Alcohol Content (BAC)	.769	.385	1.082	2.001	.053
	SE X BAC	-.239	.117	-1.571	-2.048	.048

a. Dependent Variable: Self-Disclosure

Figure 16.8 SPSS output from a moderated multiple regression estimating self-disclosure from blood alcohol content, self-esteem, and their interaction.

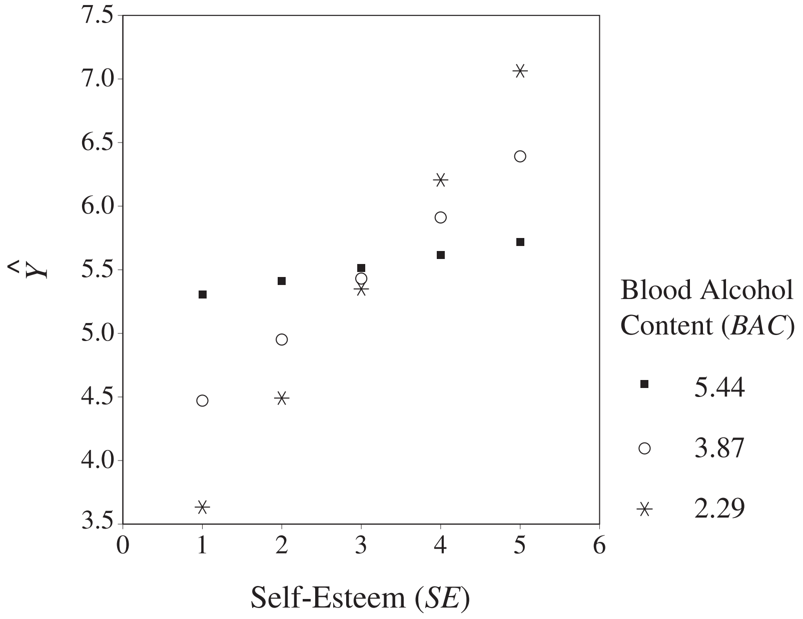


Figure 16.9 A Graphical Representation of the Interaction Between Blood Alcohol Content and Self-Esteem in Estimating Self-Disclosure

increases by one unit (because the coefficient is negative). Conversely, this interaction can be interpreted as the change in the regression weight for blood alcohol content with a one unit increase in self-esteem.

Interpretation of this interaction is made easier with a picture. Using the procedure described in the previous section, we generate estimated self-disclosure from the regression model for various values of BAC and SE. For reasons that will be clear soon, I used values of BAC equal to the mean (3.865), one standard deviation above the sample mean (5.443), and one standard deviation below the sample mean (2.287) and SE values 1 through 5 in increments of one. The estimated self-disclosure scores are then plotted in a scatterplot using different symbols for values of either SE or BAC (see Figure 16.9).[3]

The other coefficients in the regression model are interpreted just as in the previous example. The coefficient for self-esteem, b_{SE}, is the regression weight for self-esteem when $BAC = 0$. This is the relationship between self-esteem and self-disclosure if no alcohol is consumed (in which case BAC would be zero). This coefficient tells us that among a group of abstainers, two people who differ by one unit in self-esteem are expected to differ by 1.404 units in their self-disclosure, $t(36) = 2.789, p = .008$. For the same reasons as described previously, b_{BAC} has no sensible interpretation here because it is the conditional effect of blood alcohol content when $SE = 0$, but 0 is out of the range of possible values of self-esteem as measured in this study.

Probing the Interaction. We can probe this interaction using the procedure described in the previous section by estimating the conditional effect of one variable at various representative values of the other variable. To illustrate this, let's assess the relationship between self-esteem and self-disclosure at various values of blood alcohol content, using the sample mean, one standard deviation above the mean, and one

[3] The CD that comes with this book contains a couple of documents describing in detail how to generate such a plot in SPSS.

Table 16.5
Probing an Interaction in Moderated Multiple Regression

		Coeff.	s.e.	t	p
$BAC' = BAC - 2.287$					
	Constant	2.778	0.796	3.492	0.001
	Self-Esteem (SE)	0.857	0.271	3.162	0.003
	BAC'	0.769	0.385	2.001	0.053
	$SE \times BAC'$	−0.239	0.117	−2.048	0.048
$BAC' = BAC - 3.865$					
	Constant	3.992	0.567	7.045	$< .001$
	Self-Esteem (SE)	0.481	0.180	2.666	0.011
	BAC'	0.769	0.385	2.001	0.053
	$SE \times BAC'$	−0.239	0.117	−2.048	0.048
$BAC' = BAC - 5.443$					
	Constant	5.206	0.864	6.028	$< .001$
	Self-Esteem (SE)	0.104	0.243	0.426	0.672
	BAC'	0.769	0.385	2.001	0.053
	$SE \times BAC'$	−0.239	0.117	−2.048	0.048

standard deviation below the mean BAC as representative values. In the data, $\overline{BAC} = 3.865, SD = 1.578$, so one standard deviation below the mean is $3.865 - 1.578 = 2.287$ and one standard deviation above the mean is $3.865 + 1.578 = 5.443$. The results of the regressions after the necessary transformations can be found in Table 16.5. As can be seen, at one standard deviation below the mean BAC ($BAC = 2.287$) as well as at the mean ($BAC = 3.865$) the relationship between self-esteem and self-disclosure is positive and statistically significant. But at one standard deviation above the mean BAC ($BAC = 5.443$), the relationship is not significant.

16.3.3 Interaction Between a Quantitative and a Multicategorical Variable

Researchers often are interested in comparing the relationship between two variables in several naturally occurring groups or artificially created experimental conditions. For example, is the relationship between X and Y the same across all k levels of an experimental manipulation, or all k ethnic groups, $k > 2$? Or does the effect of an experimental manipulation with several different levels vary depending on the values of a second, quantitative predictor variable? When there is a multicategorical predictor variable, the same basic procedures described thus far can be applied, although there are some variations on the methodology required.

As discussed in Chapter 14, a categorical variable with more than two groups cannot be represented with a single variable in a regression model. To code group membership when there are several categories, $k - 1$ variables must be created coding group membership, where k is the number of categories. The same can be said about the required number of product terms to assess interaction between a multicategorical variable and a quantitative variable. Just as it requires $k - 1$ variables to code membership in one of k groups, it takes $k-1$ product variables to test for interaction between a categorical variable with k categories and a quantitative variable. But because it requires more than one variable to quantify and test the interaction, it is not possible to test the hypothesis of interaction with a single regression coefficient, as was possible in the analyses described previously. Instead, one must resort to hierarchical regression analysis and determine how much R^2 changes when the $k-1$ product terms coding the interaction are entered into a model without them.

To illustrate this procedure, we ask whether the relationship between exposure to political talk radio and political knowledge (Y) varies across party identification using the NES data set. First, we run a regression estimating political knowledge from political talk radio exposure and two dummy variables coding party identification. The model is:

$$\hat{Y} = 7.341 + 0.922(Talk) + 2.442(D) + 2.408(R)$$

where *Talk* is exposure to talk radio (first discussed in Chapter 13), and D and R are two dummy variables coding whether a person self-identifies as a Democrat or Republican (see section 14.4 for details). Those who identify with neither party are treated in this analysis as the reference group.

The partial regression coefficient for *Talk* is 0.922 and statistically different from zero, $t(339) = 5.125, p < 0.0005$, and the multiple correlation coefficient for this regression model is $R^2 = 0.113$. So controlling for party identification, two people who differ by one measurement unit in their exposure to political talk radio differ by 0.922 units in their political knowledge, with the person with greater exposure to political talk radio expected to have a higher political knowledge score. Combined, these three variables explain 11.3% of the variance in political knowledge. But this doesn't answer the question of interest. We want to know whether the relationship between political talk radio exposure and political knowledge depends on whether a person identifies as a Democrat, Republican, or neither of these groups. To answer this question, we ask whether a regression model that includes two additional variables that represent the interaction between political party identification and political talk radio exposure fits better than the model that excludes these interaction variables. The two variables added to the model are (a) the product of *Talk* and the dummy variable coding Democrats (D), and (b) the product of *Talk* and the dummy variable coding Republicans (R). We then estimate a regression model with the same predictors as model 1 but also including these two additional product variables. The resulting model is:

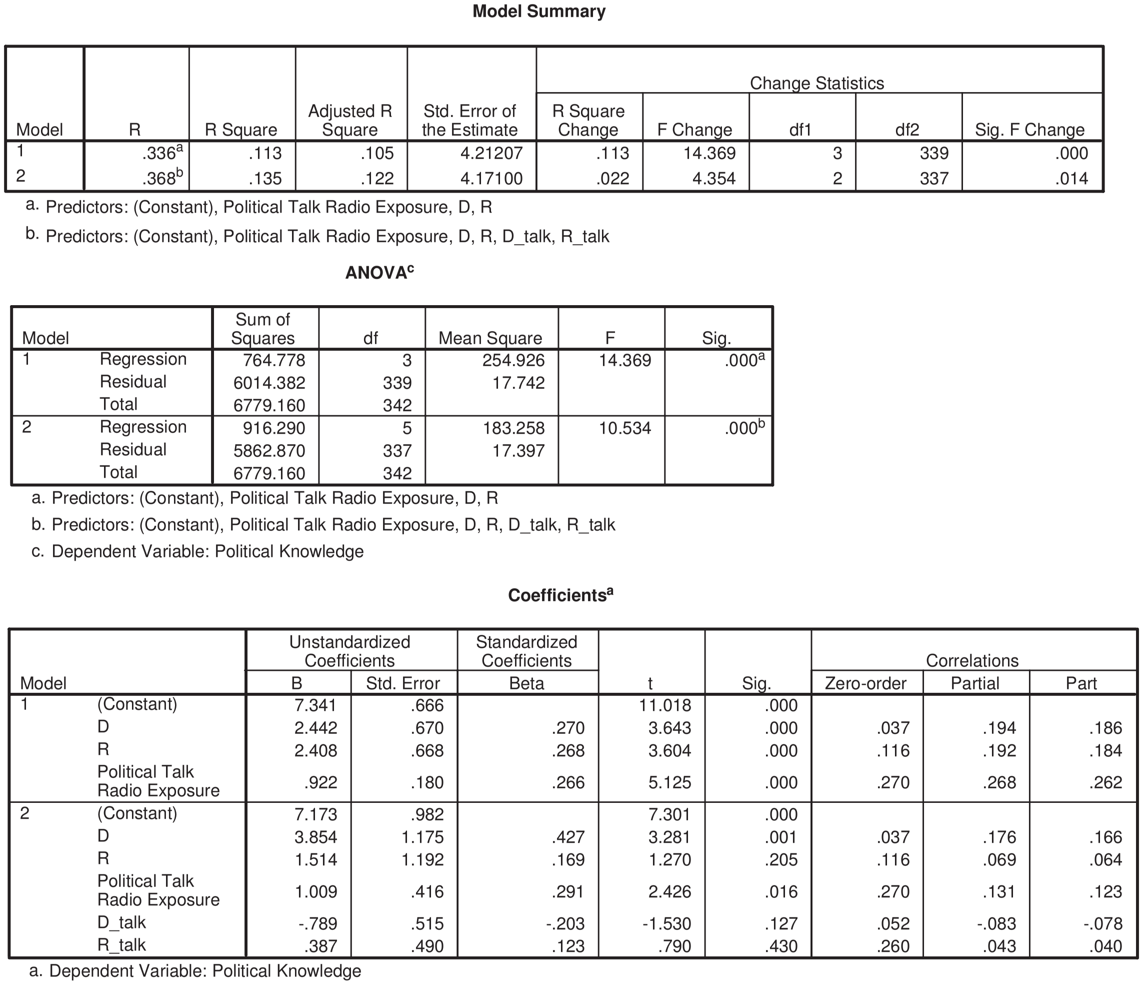
$$\hat{Y} = 7.173 + 1.009(Talk) + 3.854(D) + 1.514(R) - 0.789(Talk \times D) + 0.378(Talk \times R)$$

from the SPSS output in Figure 16.10). This model is presented graphically in Figure 16.11. The multiple correlation is $R^2 = 0.135$, so the incremental increase in variability in political knowledge explained by the addition of the two interaction terms is $\Delta R^2 = 0.135 - 0.113 = 0.022$. That is, these two variables explain an additional 2.2% of the total variance in political knowledge.

To test the null hypothesis that this increase in R^2 is attributable to a chance mechanism, we can use equation 13.14,

$$F = \frac{df_{residual} SR^2_{YX.W}}{m(1 - R^2_{YXW})}$$

To use this equation, think of variable set X as the two interaction terms $Talk \times D$ and $Talk \times R$ and variable set W as D, R, and *Talk*. In that case, ΔR^2 is the squared



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.336[a]	.113	.105	4.21207	.113	14.369	3	339	.000
2	.368[b]	.135	.122	4.17100	.022	4.354	2	337	.014

a. Predictors: (Constant), Political Talk Radio Exposure, D, R

b. Predictors: (Constant), Political Talk Radio Exposure, D, R, D_talk, R_talk

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	764.778	3	254.926	14.369	.000[a]
	Residual	6014.382	339	17.742		
	Total	6779.160	342			
2	Regression	916.290	5	183.258	10.534	.000[b]
	Residual	5862.870	337	17.397		
	Total	6779.160	342			

a. Predictors: (Constant), Political Talk Radio Exposure, D, R

b. Predictors: (Constant), Political Talk Radio Exposure, D, R, D_talk, R_talk

c. Dependent Variable: Political Knowledge

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	7.341	.666		11.018	.000			
	D	2.442	.670	.270	3.643	.000	.037	.194	.186
	R	2.408	.668	.268	3.604	.000	.116	.192	.184
	Political Talk Radio Exposure	.922	.180	.266	5.125	.000	.270	.268	.262
2	(Constant)	7.173	.982		7.301	.000			
	D	3.854	1.175	.427	3.281	.001	.037	.176	.166
	R	1.514	1.192	.169	1.270	.205	.116	.069	.064
	Political Talk Radio Exposure	1.009	.416	.291	2.426	.016	.270	.131	.123
	D_talk	-.789	.515	-.203	-1.530	.127	.052	-.083	-.078
	R_talk	.387	.490	.123	.790	.430	.260	.043	.040

a. Dependent Variable: Political Knowledge

Figure 16.10 SPSS output from a hierarchical multiple regression to estimate the interaction between political talk radio exposure and political party identification.

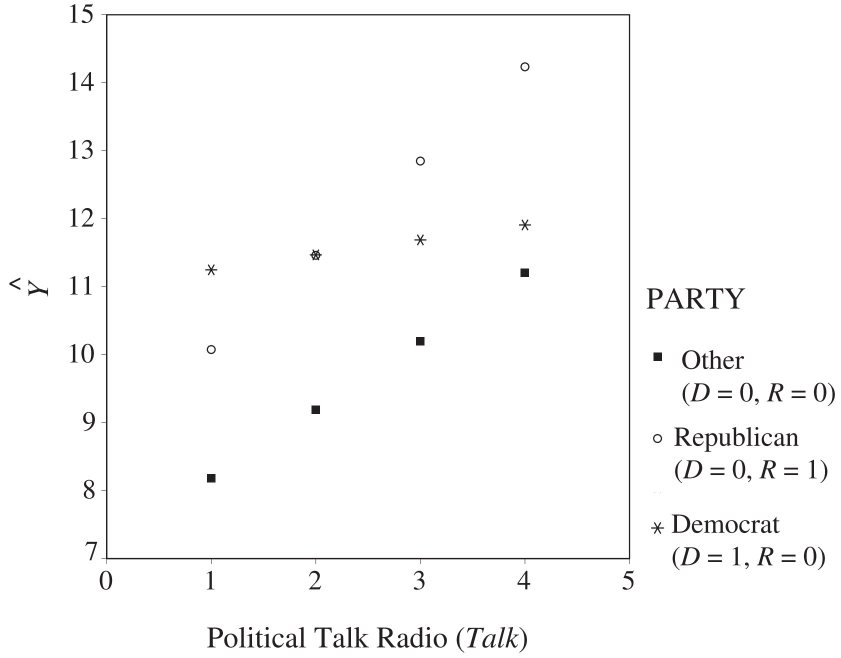


Figure 16.11 A graphical representation of the interaction between political talk radio exposure and political party self-identification in estimating political knowledge.

setwise semipartial correlation between X and Y controlling for W $(SR^2_{YX.W})$. The residual degrees of freedom for the model with the interaction terms is $df_{residual} = 337$, and $m = 2$ (the number of variables entered in the second regression model), so

$$F = \frac{337(0.022)}{2(1 - 0.135)} = 4.286$$

which is evaluated in reference to the F distribution with $df_{num} = 2$ and $df_{den} = 337$. The null hypothesis is that the relationship between political talk radio exposure and political knowledge does not vary across political identification groups. The obtained F of 4.286 exceeds the critical F for $\alpha = 0.05$ of 3.206 from Appendix D1. Using the SPSS output, we can see that the p-value for this increase is 0.014. We reject the null hypothesis and claim that the relationship between political talk radio exposure and political knowledge depends on whether a person self-identifies as a Democrat, a Republican, or some other political group.[4]

What do the Coefficients Mean? It may not be at all apparent to you how these two product variables and their corresponding regression coefficients quantify interaction. To illustrate how this is so, consider the simple regression models estimating political knowledge from political talk radio in each of the political party groups:

$$\begin{aligned}
\text{Democrats:} \quad & \hat{Y} = 11.028 + 0.220(Talk) \\
\text{Republicans:} \quad & \hat{Y} = 8.687 + 1.396(Talk) \\
\text{“Others”:} \quad & \hat{Y} = 7.173 + 1.009(Talk)
\end{aligned}$$

[4] The difference between the 4.286 computed here and from the SPSS output is simply rounding error in hand computations.

The regression coefficient for $Talk \times D$ is equal to the difference between the simple regression weight for $Talk$ in the Democrats and in the reference group. Indeed, notice that $0.220 - 1.009 = -0.789$. Similarly, the regression coefficient for $Talk \times R$ is equal to the difference between the regression weight for $Talk$ computed in the Republicans compared to the reference group: $1.396 - 1.009 = 0.387$. So these two regression coefficients do indeed quantify the differences in the regression weights. By testing the null hypothesis that both of these regression coefficients are zero, differing from each other in the sample by just chance, we are testing the null hypothesis that the relationship between political knowledge and political talk radio exposure does not vary across groups.

When an interaction term is in the model, the same rules for the interpretation of the coefficients apply. I've already discussed the interpretation of the regression coefficients for the two product terms. In the second model (model 2 in Figure 16.10), the coefficient for $Talk$ is the regression coefficient estimating political knowledge from talk radio exposure when all other variables are equal to zero. That occurs only when both D and R are equal to 0, which implies that this is the regression coefficient estimating political knowledge from talk radio exposure in the reference group. Indeed, that regression weight is 1.009, as above. The coefficient for D is the difference in estimated political knowledge between Democrats who don't listen to political talk radio (i.e., $Talk = 0$) and those who identify with neither party that don't listen to political talk radio. The coefficient for R is interpreted similarly, as the difference in estimated political knowledge between Republicans who don't listen to political talk radio and those who identify with neither party who don't listen to political talk radio.

In Chapter 15, I introduced analysis of covariance and noted in that ANCOVA assumes homogeneity of regression, meaning that the relationship between the covariate and the outcome variable is the same in all groups. You should recognize this as the assumption of no interaction between group membership and the covariate in estimating political knowledge, and now you know how to test this assumption. To do so, you follow the procedure just described.

16.3.4 Mean Centering of Predictors

It has been argued that the proper implementation of moderated multiple regression requires that the researcher first *mean center* the predictor variables prior to computing the product term representing the interaction. A variable is mean centered by subtracting the mean of the variable from each case. One argument advanced for mean centering in moderated multiple regression is that it reduces multicollinearity between the product and the constituent terms of the interaction (e.g., Aiken & West, 1991; Eveland, 1997). As discussed in Chapter 13, multicollinearity can reduce the power of significance tests in multiple regression because the variables tend to cancel each other out mathematically.

Indeed, the product of two variables X and W will tend to be highly correlated with both X and W. For instance, the correlation between self-esteem and the product of self-esteem and blood alcohol content in the example from section 16.3.2 is 0.745. And the correlation between blood alcohol content and this product is 0.846. As a result, the tolerances for the the predictors (BAC, SE, and $SE \times BAC$) are quite low—as small as 0.036 for ($SE \times BAC$). But when blood alcohol content and self-esteem are mean centered prior to computing the products, the intercorrelations are reduced substantially ($r = 0.159$ between self-esteem and $SE \times BAC$ and $r = 0.178$

Table 16.6
The Effects of Mean Centering on a Moderated Multiple Regression

		Coeff.	s.e.	t	p
Uncentered					
	Constant	1.018	1.547	0.659	0.514
	SE	1.404	0.503	2.789	0.008
	BAC	0.769	0.385	2.001	0.053
	$SE \times BAC$	−0.239	0.117	−2.048	0.048
Mean Centered					
	Constant	5.450	0.171	31.934	0.000
	SE	0.481	0.180	2.666	0.011
	BAC	0.045	0.110	0.408	0.686
	$SE \times BAC$	−0.239	0.117	−2.048	0.048
Standardized					
	Constant	5.450	0.171	31.933	0.000
	SE	0.463	0.174	2.666	0.011
	BAC	0.071	0.174	0.408	0.686
	$SE \times BAC$	−0.363	0.177	−2.048	0.048

between blood alcohol content and $SE \times BAC$), and the tolerances increase to greater than 0.85. Intuition would suggest that such mean centering would be beneficial in moderated multiple regression because it would lower the correlation between the variable representing the interaction and the two variables that define it, and this would increase the power of hypothesis tests for the regression weights.

But this is not true. Mean centering has absolutely no effect on the hypothesis test for the interaction term, as the regression weight, t statistic, and p-value will be the same regardless of whether the two predictor variables are mean centered prior to computing their product (Cohen, 1978; Cronbach, 1987; Dunlap & Kemery, 1987; Kromrey & Foster-Johnson, 1998). To illustrate, two regression models estimating self-disclosure from blood alcohol content, self-esteem, and their interaction are displayed in Table 16.6 one before mean centering and after mean centering. As you can see, the coefficient for $SE \times BAC$ is the same, as is the t statistic and p-value. Mean centering does not affect the coefficient or hypothesis test for the interaction one iota. To be sure, the regression coefficients, t statistics, and p-values for the lower order terms change as a result of mean centering, but the change is not the result of reduced multicollinearity. They change because their meaning is changed by centering, and the t statistics and p-values test a different null hypothesis.

As discussed in section 16.3.1, in a regression model of the form

$$\hat{Y} = a + b_X(X) + b_W(W) + b_{XW}(XW)$$

b_X represents the regression weight for X when $W = 0$, and b_W represents the regression weight for W when $X = 0$. But when X and W are mean centered, $X = 0$ corresponds to the mean of X and $W = 0$ corresponds to the mean of W. So b_X quantifies the regression weight for X at the mean of W, and b_W quantifies the regression weight for W at the mean of X. For example, in the regression with centered predictors in Table 16.6, b_{BAC} is the relationship between blood alcohol content and self-disclosure at the sample mean self-esteem. So two people who differ by one unit in blood alcohol content but who have a self-esteem score at the sample mean are estimated to differ by 0.045 in their self-disclosure. This difference is not statistically different from zero, $t(36) = 0.408, p = 0.686$. Similarly, b_{SE} is the relationship between self-esteem and self-disclosure at the mean blood alcohol content. Two people who differ by one unit in self-esteem but who are at the sample mean blood alcohol content are estimated to differ by 0.481 units in self-disclosure. This difference is statistically different from zero, $t(36) = 2.666, p = 0.011$.

Some argue that variables should be *standardized* prior to computing the product and estimating the moderated multiple regression model, again on the grounds that this reduces multicollinearity. True, multicollinearity is reduced when variables are standardized prior to computing the product, but as can be seen in Table 16.6, the hypothesis test for the interaction is unaffected by standardization. And although the coefficients are affected by standardization, the t statistics and p-values are not compared to when the variables are mean centered. The change in the regression coefficients has nothing to do with reduced multicollinearity. Instead, the change is the result of a difference in the unit of measurement. In regression with uncentered or centered measurements, the one-unit difference used to quantify the effect of a predictor on an outcome is a single unit in the original scale of measurement. But with standardization, one unit is *one standard deviation*. The difference between the lower-order coefficients in the regression with standardized variables compared to uncentered variables is due to the different meaning of "zero" on the measurement scales. With standardized scores, a measurement of zero is the mean, whereas with uncentered measurements, a measure of zero is just that—zero.[5]

So why all the recommendations to center or standardized predictors in moderated multiple regression? There is one reason why mean centering can be a good thing to do. In complicated models involving lots of predictors and several interactions, the tolerances can become so small that the mathematics of multiple regression explode, so to speak. Technically, the computation of a regression model requires something called *matrix inversion* in mathematics. If one predictor is very, very highly correlated with a linear combination of the other predictors, this can introduce rounding error into the computations of aspects of the regression model, yielding inaccuracies in the estimates and standard errors. That is the only sensible justification for mean centering, in my opinion. In most circumstances, it won't matter at all whether or not you mean center prior to computing the product and estimating a moderated multiple regression model. But if you choose to do so, remember that this changes the interpretation of the lower-order coefficients in the model.

[5] It is important to note that the coefficients in the moderated multiple regression model with standardized predictors are not the same as the coefficients in the *standardized regression equation* printed by most regression programs. In the latter, the product is created first, and then all the variables including the product and the outcome are standardized. In the former, only the lower order variables are standardized, after which the the product is created. The interpretation of the coefficient for the interaction in the standardized regression equation is very different than the interpretation of the interaction coefficient in the three models in Table 16.6.

16.3.5 Differences In Regression Coefficients vs. Differences in Correlations

A moderation hypothesis can also be framed as differences in correlations across groups. In the previous section, we found that the regression weight estimating political knowledge from political talk radio exposure differs between self-identifying Democrats, Republicans, and those who identify with some other political group. But perhaps the relationship between political knowledge and political talk radio exposure is stronger in one group rather than another. That is, perhaps the estimation of political knowledge from political talk radio exposure produces more accurate estimations in one group than in another. The regression weight is not a measure of strength of association in the way that a correlation coefficient is. It only quantifies how Y is estimated to change as X changes by one unit. There is a statistical test of the equality of a set of k Pearson correlation coefficients. The null hypothesis tested is that the population correlations are the same in the k groups. The alternative hypothesis is that at least two of the correlations are different.

Let's test the null hypothesis that the correlation between political talk radio and political knowledge is the same across these three groups. In the NES data, the sample correlations are $r = 0.059$, $r = 0.439$, and $r = 0.291$ in the 141 Democrats, 147 Republicans, and 55 respondents who identify as neither, respectively. To conduct this test, the difference between the sample correlations is converted to a chi-squared statistic with equation 16.7:

$$\boxed{\chi^2 = \sum(n_j - 3)(Z_{r_j} - \overline{Z}_r)^2} \tag{16.7}$$

where Z_{r_j} is the correlation between X and Y after Fisher's r-to-Z transformation (equation 12.19), n_j is the sample size in group j, and $\overline{Z}_r$ is the weighted mean of the k Fisher-transformed correlations, defined as

$$\boxed{\overline{Z}_r = \frac{\sum(n_j - 3)Z_{r_j}}{n - 3k}} \tag{16.8}$$

where n is the total sample size.

The conversion of r to Fisher's Z using equation 12.19 yields Z_r values of 0.059, 0.471, and 0.300 for Democrats, Republicans, and those who identify as neither, respectively. From equation 16.8,

$$\overline{Z}_r = \frac{(138)0.059 + (144)0.471 + (52)0.300}{343 - 9} = 0.274$$

Applying equation 16.7 yields

$$\chi^2 = (138)(0.059 - 0.274)^2 + (144)(0.471 - 0.274)^2 + (52)(0.300 - 0.274)^2 = 12.002$$

The p-value for $\chi^2 = 12.002$ assuming the null hypothesis is true can be derived from the χ^2 distribution with $(k - 1)$ degrees of freedom. From Appendix C, the critical χ^2 for $df = 2$ and $\alpha = .05$ is 5.991. The obtained χ^2 statistic does exceed this critical value, so the null hypothesis is rejected, $p <$ 05. Indeed, the obtained χ^2 exceeds the critical value for $\alpha = 0.01$, so the p-value is less than 0.01. So there is evidence that the correlation between political knowledge and exposure to political talk radio differs as a function of political party identification.

Although this test will tend to produce the same substantive conclusion about interaction as will moderated multiple regression, they can conflict. Which you use depends on the question of interest. The approach of comparing correlations asks whether the relationship between two variables is equally strong across the k groups, whereas the moderated multiple regression approach determines whether a one unit increase in a predictor is associated with the same expected difference in Y across all k groups. Although there are occasions where you might be interested in comparing the correlations, the problem with this approach is that the correlation between two variables can vary across groups if the range or variances of either the predictor or outcome varies substantially across groups.

This procedure should be used only if the correlations being compared are statistically independent. A necessary but not sufficient condition for correlations to be independent is that each unit must provide data to the computation of only one of the correlations involved in the comparison. This criterion is met in this example because the each person provides data to only the correlation between political discussion and knowledge in that person's group. This procedure could not be used to compare, for example, the correlation between political knowledge and political discussion and between political knowledge and newspaper exposure. In that case, each unit would be providing data to both of the correlations being compared. So they cannot be considered statistically independent. Correlations can also be nonindependent if units in the data are paired in some fashion, such as husband and wife couples. In this example, there is no pairing between units in the sample, so we are safe in using this procedure. Statistical procedures for comparing nonindependent correlations are described by Griffin and Gonzales (1995), Meng, Rosenthal, and Rubin (1992), Raghunathan, Rosenthal, and Rubin (1996), and Steiger (1980).

16.3.6 Statistical Control of Covariates

In moderated multiple regression, one or more covariates can be statistically controlled simply by including them in the regression model. The regression coefficient for the interaction in a moderated multiple regression model then quantifies the interaction controlling for the covariates. For example, in the analysis reported in section 16.3.2 we may have wanted to control for a participant's weight when assessing the interaction between blood alcohol content and self-esteem. Had weight been measured, it could have simply been included in the regression model, and all interpretations would be based on the statistical control of individual differences in body weight.

16.3.7 Required Terms in a Moderated Multiple Regression Model

In a regression model with an interaction between variables X and W, is it necessary to always include both X and W in the model, or can they be deleted if not statistically significant? The answers to these questions, respectively, are "yes" and "no." In order for the various terms in a moderated regression model to be estimated correctly and interpreted as described here, it is necessary that lower order variables that define the interaction be included in the regression model, regardless of whether or not their regression coefficients are statistically significant. A failure to do so will produce largely meaningless results that cannot be interpreted as described here. For this reason, you should *never* use stepwise variable entry (see section 13.6.4) to build a model that contains interactions because the variable entry algorithms used by stepwise procedures will not recognize this important constraint.

Although one should always include the lower order variables contributing to an interaction whenever the product of those variables is in a model, the same cannot be said about retaining a nonsignificant interaction term. Because the presence of an interaction term in a moderated multiple regression model changes the interpretation of the lower order coefficients from partial regression weights to conditional regression weights, it is a good idea to include product terms in a final model only if the interaction is statistically significant. If the interaction term is not significant, estimate a new regression model without it.

As I have illustrated throughout this section, the presence of an interaction term in a regression model changes the interpretation and tests of significance of the regression coefficients for the lower order variables that define the interaction. For this reason, hierarchical entry should be used if one is interested in first assessing the partial relationships between the predictor and outcome variables prior to assessing interaction. In the first stage of a hierarchical model, the predictor variables of interest are used to predict the outcome. At a second stage, the interaction is then added. If the interaction is nonsignificant, then all discussion of the results should be based on the first stage of the regression. If the interaction is significant, the coefficients for the lower order terms in the second stage model should be interpreted as conditional regression weights (which may or may not be meaningful), and the researcher can then probe the interaction using the methods described in this chapter.

16.3.8 The Pitfalls of Subgroup Analyses to Test Moderation

One strategy you will see in the communication literature for testing moderation hypotheses is to estimate a regression model several times, once for each of two or more subgroups in a sample, and then descriptively compare either the standardized or unstandardized regression coefficients for each variable across the groups. According to this approach, if one of the regression coefficients representing the partial relationship between a predictor and an outcome appears to differ across groups, then the grouping variable is considered a moderator of the relationship between that predictor and the outcome. For example, if predictor variable i is statistically significant in one group but not in another group, the grouping variable is deemed a moderator of the relationship between predictor i and the outcome.

There are two major problems with this strategy. First, a descriptive difference between regression coefficients (either standardized or unstandardized) for variable i across regression models does not imply that the relationship (simple or partial) between predictor i and the outcome variable differs across groups. Indeed, you'd expect regression coefficients from models estimated in different groups to differ from each other as a result of sampling error. Two coefficients may differ descriptively but not differ statistically (i.e., by more than "chance"). And evidence that the relationship between variable i and the outcome (again, simple or partial) is statistically significant in one group but not the other cannot be used as evidence of moderation when the groups differ in sample size, because the size of the standard error is determined in part by the size of the sample. And given that the standard error of a partial regression weight is determined in part by intercorrelations between predictors (see section 13.6.3), differences across groups in the predictor variable intercorrelations can also affect whether a variable is statistically significant in one group or the other. So sample size, predictor variable intercorrelations, and statistical significance are confounded.

Second, in spite of recommendations that standardized regression coefficients be routinely reported and used when comparing models across groups or studies (e.g., Hunter & Hamilton, 2002), standardized regression coefficients estimated in different subgroups are often not comparable. If the variance of either the predictor or the outcome variable differs across groups, standardized regression coefficients are expected to differ from each other even if predictor i has the same effect on Y across groups. Even minor differences in these variances across groups can produce differences in standardized coefficients. But such variations in variance across groups will have little to no effect on unstandardized coefficients (see, e.g., Blalock, 1967; Linn & Werts, 1969).

It is generally accepted in the field of statistics (even if not widely practiced in communication) that comparisons of regression coefficients for variable i between subgroups in a sample should be based on *unstandardized* coefficients, and a formal test of the significance of the difference using moderated multiple regression (or an equivalent strategy) should be conducted and the null hypothesis rejected before one can claim that a predictor variable's effect on an outcome variable differs across groups. Subgroup regression analyses, especially when based on standardized regression coefficients, are not informative about whether a variable's effect on an outcome differs across subgroups of the sample. For a good and very readable discussion on the problems of subgroup analysis to assess moderation hypotheses, see Newsom, Prigerson, Schultz, and Reynolds (2003).

16.4 Simplifying the Hunt for Interactions

In a moderated multiple regression with k predictors, there are "k choose 2" possible interactions between two predictors. For example, with 4 predictor variables, there are $4(3)/2 = 6$ possible 2-way interactions, with 5 predictors that are $5(4)/2 = 10$ possible 2-way interactions, and with 10 predictors there are $10(9)/2 = 45$ possible interactions between two predictors. Given the large number of possible interactions between two variables in even relatively simple multiple regression models, is it necessary to test for all of them? Some of them? If only some of them, which ones? In this section, I discuss some strategies for thinking about how to manage tests for interactions in linear models such as regression and analysis of variance.

The first question to address is whether one needs to bother with testing for interactions in the first place. It is clear from the communication literature that one common strategy is to assume that interactions don't exist. On the surface this might seem silly, but in fact there is some justification for ignoring the possibility that predictors may interact. There is no obligation that you test for interaction between predictors just because it is possible to do so. A failure to include interactions that should be in a regression model will of course lead to a regression model that is at best an oversimplification or, at worst, just plain wrong. But all regression models are wrong in some sense. For example, you might argue from the results of a regression analysis that because the relationship between X and Y persists even after controlling for W and Q that the relationship is not spurious. But there is an infinite number of other predictors that you could have included in the regression model but simply did not, either because you didn't measure the variables or you simply didn't think of including them in the model. If you had included them, the relationship between

X and Y may have disappeared. A regression model that fails to include potentially important predictors is called an *underspecified model.* The communication literature is filled with underspecified regression models, because we can never know for certain which potentially important variables have not been included in the model. Whether a model is underspecified or not is part a statistical judgment but also a theoretical judgment. A critic of your research may argue that you failed to control for something important. Such a criticism is usually a principled argument, in that the critic believes that there is a specific variable that you should have controlled for and that, had you done so, you would have ended up with a dramatically different result. By the same reasoning, a critic could make the case that a failure to include an interaction may lead to a misleading result, but such a criticism is usually principled, in that the critic typically would have particular reason for believing that there is an interaction that you should have included in your regression model.

My point is you should test for interaction if there is a principled argument leading you to expect there to be an interaction between two specific variables in the regression model. The primary principled argument for testing for an interaction is that your hypothesis or the theory you are testing postulates that an interaction between two of the variables should exist. If you are testing a theory or hypothesis that predicts an interaction, you darn well better test for it. If you don't test for it, you aren't testing the theoretical proposition correctly (c.f., Eveland, 1997). But if there is no reason to expect an interaction, and you can conceive of no principled arguments that a critic might make for why any of the variables should interact, then you have no obligation to test for interaction.

Nevertheless, the possibility of missing an interaction should loom large in your thinking. There is little harm in testing for interaction even if you don't have an a priori reason to believe such an interaction might exist. It is exciting to discover something unexpected, and unexpected discoveries can often lead to new research questions, new ways of thinking about old questions, or can even revolutionize and move theory in directions it otherwise wouldn't have gone. So I encourage you to explore your data in search of interactions for no reason other than the possibility that the unexpected may appear in your data. But how do you manage so many possible statistical tests given the number of possible interactions in even relatively small regression models?

One strategy is to test for interactions as a set. For example, in a regression model with 4 predictors, you might add all six possible interactions at a second step and see if R^2 increases significantly. If so, this suggests that there is at least one interaction between two of the predictors, and hopefully the coefficients for the interaction and tests of significance for each will tell you which interactions are significant in the set. Or you might define sensible subsets of possible interactions that are worth testing as a set. For example, if X is an experimental manipulation and the primary variable of interest in your study and W, Z, and Q are three covariates, you could see if adding $X \times W$, $X \times Z$, and $X \times Q$ as a set increases R^2 significantly. If not, this suggests no interaction between the experimental manipulation and any of the covariates. Having determined that X doesn't interact with W, Z, or Q, you could then test whether there are any interactions between W, Z, and Q by seeing if adding $W \times Z$, $W \times Q$, and $Z \times Q$ significantly increases R^2 compared to the model that includes only X, W, Z, and Q by themselves. If not, then stop searching and conclude that there are no interactions. If the increase in R^2 is statistically significant, look at the individual coefficients and their tests of significance. Another possibility is to test each interaction separately. So if there are 4 predictors, you run 6 additional regressions including one of the possible interactions in each regression model.

Of course, by fishing around in your data for statistically significant relationships, you are bound to find a small p-value now and then that reflects nothing other than sampling error or "chance." Some kind of multiple test correction (such as a Bonferroni correction) is justified when hunting in your data for something statistically significant worth reporting. Alternatively, repeat the study with a new set of participants to see if the interaction you found when mining the first data set replicates in the new data.

In factorial research designs, the default approach is to include both the main effects and the interactions in the ANOVA. This habit no doubt stems from where communication scientists have learned about the analysis of experiments. But the same logic I discussed above applies to factorial ANOVA. There is no obligation to test for interaction just because it is possible to do so. You should test for interaction if you have a reason to do so (such as your hypothesis predicts an interaction) or if you are just curious. If an interaction in a factorial ANOVA is not statistically significant, a strong argument can be made for eliminating it from the analyses (most good statistics programs have options for excluding interactions from an analysis of variance). If an interaction is not statistically significant, then the more parsimonious and better fitting model of the data is one that excludes the interaction term. And in unbalanced designs, the interaction is likely to be correlated with the main effects, so including the interaction when nonsignificant can lower the power of the F tests for the factors by reducing the sum of squares for the effect of interest and therefore the mean square for that effect. Conversely, including a nonsignificant interaction term can artificially lower MS_{error}, increasing the probability of a Type I error in tests for the factors.

16.5 Why Not to Categorize Prior to Testing for Interaction

In section 16.3, I introduced moderated multiple regression as a means of assessing whether the relationship between two variables X and Y varies as a function of a potential moderating variable W. Although moderated multiple regression is the preferred method for testing for interactions involving a quantitative predictor variable, unfortunately this procedure is not often used when it should be. A common strategy that you will see in the communication literature is for the researcher to take one of more of the quantitative predictor variables and place people into categories based on their scores on the quantitative predictor prior to testing for interaction using a factorial ANOVA. Such categorization of quantitative variables can take many forms, the most common being dichotomization through a median or mean split, where the investigator creates "low" and "high" groups prior to data analysis based on whether participants score below or above the sample median or mean on some quantitative measure. Other forms of categorization include trichotomization (the creation of three groups), bivariate group construction, where an investigator creates a special category of participants that exceed some criterion on more than one variable (e.g., classifying people into a group based on whether the person is above the median on 2 different measures), or arbitrary categorization, where the groups are defined based on whether a participant scores higher or lower than some arbitrary value other than the median or mean. For example, a researcher may classify participants as knowledgeable or unknowledgeable about a political candidate based on whether he or she can correctly answer 50% or more of the questions in a set about the candidate.

Many arguments have been presented against categorization of quantitative variables, although these arguments tend to focus on relatively simple analyses, such as testing for association or comparing two groups (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002). My interest in this chapter is tests of interaction, and in this section I try to make the case as to why such categorization should be avoided. But some of the arguments in the literature focused on simpler analyses are relevant, so I use them when necessary.

16.5.1 Classification Errors

One way of thinking about the damage to analyses caused by categorization is to ask how frequently participants are likely to be classified into the wrong group using a categorization procedure. Remember from Chapter 6 that the observed scores resulting from measurement are used as proxies for the true scores. Unless your measurement procedure is perfectly reliable, the observed scores on the variable being used to produce the categories are not going to be equal to the true scores of what is being measured. Our measurement procedures are not perfect. Measurements will almost always contain some random error, meaning that the match between each case's observed score and their true score is not going to be exact. So, for example, somebody whose observed score is below the sample median on some kind of individual difference variable such as communication apprehension may actually be above the median if his or her true score could be known exactly. Of course, the true scores are usually unknown, and so the observed scores are used to derive category membership. Random measurement error means that some participants that really are above (or below) the median on the true score may be misclassified as low (or high) on the variable being measured, with the frequency of misclassification depending on how reliably the construct is measured.

Table 16.7 illustrates the effect of measurement error on classification accuracy using either a median split or a trichotomization procedure (based on dividing the participants into "low," "middle," and "high" groups as a function of whether their observed scores are in the lower, middle, or upper third of the distribution). The numbers in this table were generated through a simulation and assume that the true scores and errors in measurement are normally distributed. For example, using a measurement

Table 16.7
Estimated Percent of Cases Misclassified As a Result of Categorization

Reliability of Measurement	Median Split	Trichotomization	2 × 2 Cross Class-Classification
1.00	0	0	0
0.90	10	18	19
0.80	15	26	27
0.70	18	32	33
0.60	22	37	39
0.50	25	41	44

procedure with minimally acceptable reliability to generate groups (generally, 0.70 is used by communication researchers), roughly 18% of participants are likely to be misclassified using a median split on the observed scores or 32% using a trichotomization procedure. Of course, the problem is less severe when the original measurements are more reliable, but even then the problem is not trivial. Using a second quantitative variable also measured with some error to produce a 2×2 classification of participants (as either above/below the median on one variable and above/below the median on the other) exacerbates the problem further. The last column in Table 16.7 provides expected misclassification assuming both variables are measured with the same reliability. For example, if both measures have reliability of 0.70, roughly 34% of participants will be misclassified in a 2×2 table based on a median split of each variable.

To put these numbers into a meaningful context, consider an investigator studying how men and women differ in communication apprehension. The misclassification estimates in Table 16.7 can be thought of as equivalent to the effect of misidentifying a person's sex in the data set. It would be potentially disastrous if a researcher of sex differences mistakenly misidentified the sex of 20% or more of his or her participants. Indeed, a researcher who later discovered such an error after publishing the results would likely feel an obligation to publish a correction of some sort. Yet categorization of quantitative variables produces such miscodings routinely and with certainty whenever a variable is measured with error.

16.5.2 Smaller Effect Sizes and Reduced Statistical Power

Categorization of a quantitative predictor prior to assessing interaction involving that variable tends to lower statistical power of tests of interaction as well as reduce the size of interaction effects. Remember that power is the ability of a hypothesis test to reject a false null hypothesis. Higher power is better.

To illustrate the effects of categorization on power and effect size, I present results of a small simulation in Table 16.8. To conduct this simulation, I generated samples of various sizes from two hypothetical populations with different relationships between an outcome variable and a quantitative predictor. In one population, the relationship between X_1 and Y was defined as $Y = \beta_1(X_1) + e$, where X_1 and e were random standard normal variables. This procedure was repeated but sampling from a different population, using not β_1 but β_2, where β_2 was set to a value different from β_1. In half of the simulations, X_1 was then dichotomized using a median split based on the sample median computed after combining both samples, creating "high" and "low" groups on X_1. X_1 was then recoded $X_1 = 0$ for the low group, and $X_1 = 1$ for the high group. The presence of interaction was then tested using a 2 (population sampled) × 2 (high vs. low on X_1) ANOVA. The proportion of the total variance in Y attributable uniquely to the interaction was computed, as was a test of significance for the interaction, using $\alpha = .05$ for rejection of the null hypothesis. In the other half of the simulations, these same statistics were computed testing for interaction using moderated multiple regression. This procedure was repeated 10,000 times for various combinations of β_1, β_2 and n (sample size in each group).

As can be seen in Table 16.8, both the proportion of variance in Y uniquely attributable to the interaction (η^2) and the power of the hypothesis test for interaction were lower for the ANOVA strategy compared to moderated multiple regression (although as you would expect, increasing the sample size reduced the differences in power because power converges to 1 for both tests with increasing sample size). Thus, the

Table 16.8
Comparing Factorial ANOVA to Moderated Multiple Regression

			Mean Interaction η^2		Power	
β_1	β_2	n per group	ANOVA	MMR	ANOVA	MMR
0	0.7	20	0.082	0.109	0.361	0.530
		50	0.070	0.102	0.748	0.913
		100	0.067	0.100	0.966	0.997
		250	0.064	0.099	1.000	1.000
0.3	0.5	20	0.029	0.030	0.080	0.090
		50	0.015	0.017	0.119	0.165
		100	0.010	0.012	0.192	0.282
		250	0.007	0.010	0.406	0.603
0.3	−0.3	20	0.071	0.096	0.285	0.411
		50	0.059	0.088	0.624	0.819
		100	0.056	0.085	0.908	0.984
		250	0.054	0.083	1.000	1.000
0.3	0.7	20	0.042	0.048	0.157	0.218
		50	0.029	0.038	0.330	0.493
		100	0.024	0.034	0.574	0.792
		250	0.021	0.032	0.924	0.993
0.7	−0.7	20	0.212	0.314	0.857	0.973
		50	0.210	0.321	0.999	1.000
		100	0.209	0.325	1.000	1.000
		250	0.210	0.327	1.000	1.000

categorization-followed-by-ANOVA strategy tended to produce smaller effect size estimates and was less likely to reject the false null hypothesis of no interaction when interaction was present than did moderated multiple regression.[6]

These results illustrate the lower statistical power and effect size estimates that result from categorization of quantitative variables prior to assessing interaction. Categorization of quantitative variables followed by factorial ANOVA reduces an investigator's ability to detect interactions when they are present compared to moderated multiple regression (e.g., Aiken & West, 1991, pg. 167–168; Bissonnette, Ickes, Berstein, & Knowles, 1990).

[6] The overall decline in the interaction η^2 as a function of sample size seen in Table 16.8 is attributable to the fact that η^2 is an upwardly biased estimate of effect size, but the bias decreases as sample size increases.

16.5.3 Spurious Statistical Significance

The skeptical reader may have another argument in favor of categorization followed by ANOVA. The power of a statistical test is irrelevant if one has successfully rejected a null hypothesis, so what harm is there in categorizing and testing for interaction using ANOVA if one has indeed found a number of interpretable effects after doing so? My argument thus far is that categorization of quantitative variables increases the probability of Type II errors (failing to reject a false null hypothesis). So why worry about failure to reject a null hypothesis if you have already done so successfully in spite of the problems with categorization?

In some circumstances, categorization of a quantitative variable can actual increase the likelihood of falsely rejecting a true null hypothesis and claiming support for a hypothesis or theory that is in fact false. That is, categorization of quantitative variables prior to analysis can yield spuriously significant effects. The statistical evidence is highly technical and summarized in a variety of sources. The most well known argument applies to non-experimental designs, where two quantitative variables are both dichotomized and interaction is tested with a 2 × 2 ANOVA. Maxwell and Delaney (1993) show that when two correlated variables are both dichotomized prior to analysis with a factorial ANOVA, the probability of a significant main effect can be much higher than the level of significance chosen for the test. In other words, the p-value for one of the main effects can be substantially underestimated, increasing the likelihood of a Type I error. If one of the dichotomous variables is an experimental manipulation and participants are randomly assigned to conditions, this is less likely to happen in a simple 2 × 2 design because random assignment would result in a zero or near zero correlation between the quantitative independent variable (in either the original or dichotomized form) and the levels of the experimental manipulation. But in more complicated analyses (e.g., a 2 × 2 × 2 design), the presence in the analysis of any interaction involving the two dichotomized variables can produce spuriously significant effects. Dichotomization of variables and conducting an ANOVA can also yield a spuriously significant interaction if the true relationship between one of the IVs and the DV is curvilinear (Maxwell & Delaney, 1993). If an predictor variable is dichotomized prior to analysis, it is impossible to assess whether a curvilinear relationship exist between that variable and dependent variable, and any curvilinearity that does exist (i.e., prior to dichotomization) can show up as interaction.

16.5.4 Artifactual Failures to Replicate Findings

Anyone familiar with the communication literature knows that research findings are anything but consistent. In seemingly similar studies, one investigator may report one finding, and a different investigator might report something completely different. There are many different reasons an investigator may fail to replicate previous findings, among them being different populations studied, important between-study variations in the stimuli, differences in sample size, or simply time passing or society changing in an important way relevant to the phenomenon being studied. It is also true that two investigators who categorize quantitative variables prior to analysis may get very different results even though the same basic relationship between the variables exists in the data (c.f., Hirsch, 1980; Hunter & Schmidt, 1990; Sedney, 1981; Viele, 1988). Figure 16.12 graphically represents the relationship between a quantitative measure X (perhaps an individual difference like extroversion or a behavioral measure such as the number of hours spent watching television on a typical day) and some dependent

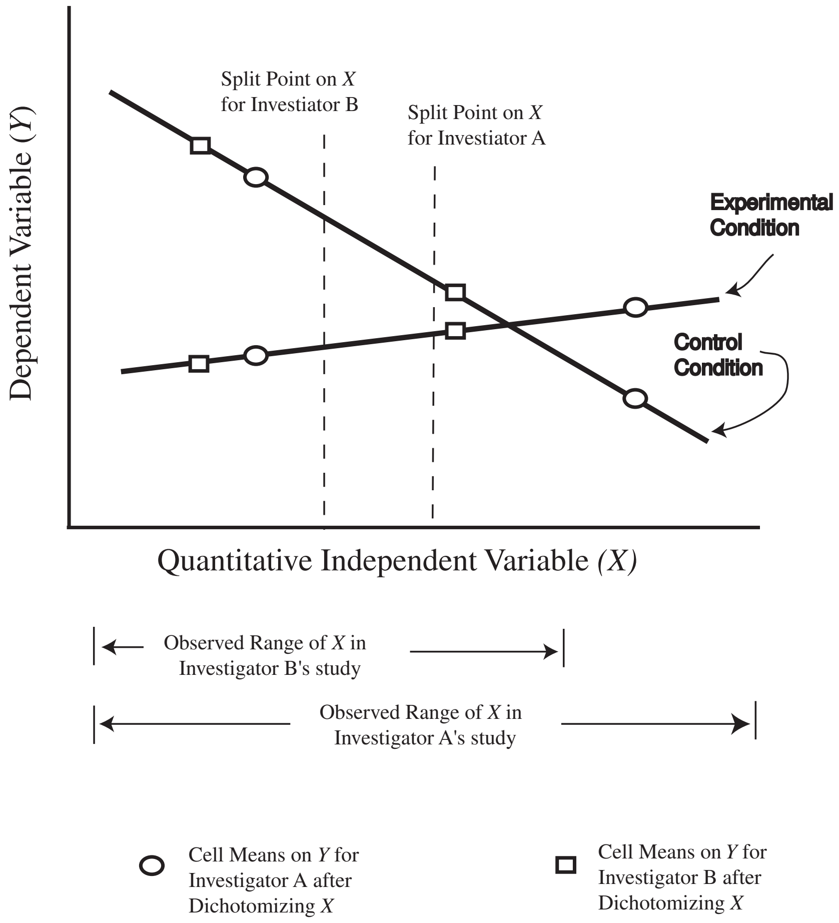


Figure 16.12 The effects of different median split points on study outcomes.

variable Y in a hypothetical experimental study. As can be seen, for participants in a control group, the relationship between X and Y is negative. But the relationship between X and Y is close to zero or perhaps positive in the experimental condition. Investigator A has a sample of participants representing a wide range of values on X. Investigator B, by contrast, has a sample that is restricted in the range of X, with the bulk of the participants being in the low to middle range. Both investigators are interested in the interaction between X and experimental treatment vs. control and examine this by dichotomizing the participants at the median on X and analyzing the study with a 2×2 ANOVA. As represented in Figure 16.12, investigator A's sample median is much higher than investigator B's sample median. Notice that investigator B's "high" group is not high at all relative to investigator A's sample. Indeed, what investigator B calls "high" on X investigator A would consider "moderate" or about average. In Figure 16.12, the circles and squares represent the cell means in this 2×2 design for investigator A and B, respectively. Investigator A would likely report an interaction between experimental condition vs. control and X and perhaps a small main effect of X. A simple effects analysis might show that among those low on X, the control condition had a higher mean on the dependent variable compared to the experimental condition. But among those high on X, exactly the opposite effect occurred, with a larger mean in the experimental condition. In contrast, Investigator

B would likely report a main effect of experimental condition as well as an interaction between experimental condition and X. An analysis of simple effects might yield the finding that differences between the experimental and control group are much larger in the low group than in the high group (and thus the interaction), but the direction of the difference is the same.

So the apparent conflict in research findings between two studies can be an artifact of where on the quantitative variable two investigators split the sample. In this example, this problem might have been detected by one of the investigators if he or she knew prior to interpretation that samples in the two studies differed widely in their representation of the range of scores on X. But rarely do researchers have such intimate familiarity with the data of other researchers, so in most circumstances neither investigator would detect this problem (nor would journal reviewers or editors) and the result would be a conflicting and unnecessarily confusing literature. This might motivate some new investigator C to seek out some important difference between the methodologies of the two studies in attempt to design a study to explain the discrepancy in the hopes of advancing theory. Clearly, such attempts would be in vain, as there is no real discrepancy in the pattern of relationships between the individual difference and the dependent variable across the studies.

16.5.5 Is Categorization Ever Sensible?

Is there any reason why it might be sensible to categorize prior to analysis? There are two. First, categorizing is sensible when true categories exist and observed individual differences other than those attributable to category membership can be construed as measurement error. Second, it is sensible to categorize if there is a qualitative difference that results in a shift from one measurement to the next. For instance, if you asked people how many cigarettes they smoke each day, it might be sensible to categorize people into nonsmoking (zero cigarettes) and smoking (one or more cigarettes) groups if you were studying the effect of *any* smoking (rather than *how much*) on some outcome variable or as a moderator. But this wouldn't be sensible if you were interested in how small variations between people in the number of cigarettes smoked related to an outcome. Otherwise, in general, one should not categorize unless a convincing argument (as opposed to just an assumption) can be presented that categorization produces more meaningful measurement as it relates to the purpose of the research than does the use of the original measurements (c.f., Cohen, 1983; MacCallum et al., 2002).

16.6 Summary

Although the concept of interaction is relative simple conceptually, it can be tricky to test statistically. There are many forms that interaction can take, and may different statistical approaches to testing for interaction. We have only scratched the surface of the topic, and I encourage you to consult more advanced books referenced in this chapter for guidance on other forms of interaction and how to test for such forms of interaction in your data.

When the predictors are all categorical, the standard approach to testing for interaction is factorial analysis of variance. Although all good statistical programs can conduct a factorial ANOVA it is important to understand the interpretational differences that result when a design is unbalanced compared to when it is balanced. Most importantly, main effects are tests of differences between unweighted marginal means, not weighted marginal means, and so you can't just pretend that a variable in a factorial design doesn't exist when generating the means for your interpretation. If your statistics program doesn't automatically generate the unweighted marginal means, you need to calculate each of the cell means and then derive the unweighted means yourself before interpreting the main effects.

When one of the predictors presumed to interact with another predictor is quantitative, moderated multiple regression is the strategy of choice. Moderated multiple regression is used to test whether a regression coefficient (or a partial regression coefficient if there are covariates in the model) varies systematically as a function of variations in a second predictor variable. The inclusion of interaction terms in a regression model drastically alters the interpretation of the regression coefficients for variables that constitute the interaction. Rather than measures of partial association, those coefficients become measures of conditional association.

Although this may be the most difficult chapter in the book to read and master, your effort was well worth the effort. Understanding how interaction is conceptualized theoretically and tested statistically will take you a long way in life as a reader and producer of communication science.