

Testing Moderator and Mediator Effects in Counseling Psychology Research

Patricia A. Frazier
University of Minnesota

Andrew P. Tix
Augsburg College

Kenneth E. Barron
James Madison University

The goals of this article are to (a) describe differences between moderator and mediator effects; (b) provide nontechnical descriptions of how to examine each type of effect, including study design, analysis, and interpretation of results; (c) demonstrate how to analyze each type of effect; and (d) provide suggestions for further reading. The authors focus on the use of multiple regression because it is an accessible data-analytic technique contained in major statistical packages. When appropriate, they also note limitations of using regression to detect moderator and mediator effects and describe alternative procedures, particularly structural equation modeling. Finally, to illustrate areas of confusion in counseling psychology research, they review research testing moderation and mediation that was published in the *Journal of Counseling Psychology* during 2001.

If you ask students or colleagues to describe the differences between moderator and mediator effects in counseling psychology research, their eyes are likely to glaze over. Confusion over the meaning of, and differences between, these terms is evident in counseling psychology research as well as research in other areas of psychology (Baron & Kenny, 1986; Holmbeck, 1997; James & Brett, 1984). This is unfortunate, because both types of effects hold much potential for furthering our understanding of a variety of psychological phenomena of interest to counseling psychologists.

Given this, our goals here are to (a) describe differences between moderator and mediator effects; (b) provide nontechnical, step-by-step descriptions of how to examine each type of effect, including issues related to study design, analysis, and interpretation of results; (c) demonstrate how to analyze each type of effect through the use of detailed examples; and (d) provide suggestions and references for further reading. We focus on the use of multiple regression because it is an accessible data-analytic technique contained in major statistical packages that can be used to examine both moderator and mediator effects (Aiken & West, 1991; Baron & Kenny, 1986; Cohen, Cohen, West, & Aiken, 2003; Jaccard, Turrisi, & Wan, 1990). When appropriate, however, we also note

limitations of using multiple regression to detect moderator and mediator effects and describe alternative procedures, particularly structural equation modeling (SEM). In addition, to illustrate areas of confusion in counseling psychology research, we review research testing moderation and mediation that was published in the *Journal of Counseling Psychology (JCP)* during 2001. Finally, we want to stress that, although our goal was to summarize information on current best practices in analyzing moderator and mediator effects, we strongly encourage readers to consult the primary sources we reference (and new sources as they emerge) to gain a better understanding of the issues involved in conducting such tests.

DIFFERENCES BETWEEN MODERATOR AND MEDIATOR EFFECTS

Consider, for a moment, your primary area of research interest. More than likely, whatever domain you identify includes research questions of the form “Does variable *X* predict or cause variable *Y*?”¹ Clearly, questions of this form are foundational to counseling psychology. Examples include correlational questions such as “What client factors are related to counseling outcomes?” as well as causal questions such as “Does a certain counseling intervention (e.g., cognitive therapy) increase well-being?” (see Figure 1A for a diagram). However, to advance counseling theory, research, and practice, it is important to move beyond these basic questions. One

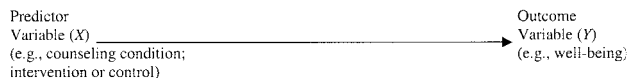
Patricia A. Frazier, Department of Psychology, University of Minnesota; Andrew P. Tix, Department of Psychology, Augsburg College; Kenneth E. Barron, Department of Psychology, James Madison University.

We thank Michele Kielty Briggs, Bryan Dik, Richard Lee, Heather Mortensen, Jason Steward, Ty Tashiro, and Missy West for their comments on an earlier version of this article and David Herring for his assistance with the simulated data.

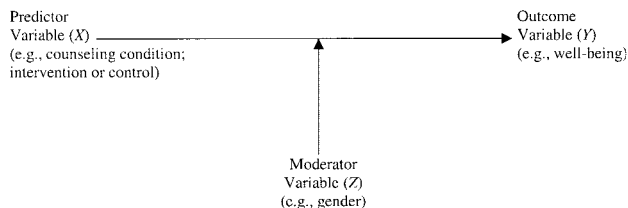
Correspondence concerning this article should be addressed to Patricia A. Frazier, Department of Psychology, University of Minnesota, N218 Elliott Hall, 75 East River Road, Minneapolis, MN 55455. E-mail: pfraz@umn.edu

¹ For the sake of simplicity, we generally use the term *predictor variable* to refer to both a predictor variable in correlational research and an independent variable in experimental research. Likewise, we generally use the term *outcome variable* to refer to both an outcome variable in correlational research and a dependent variable in experimental research.

A. Direct Effects



B. Moderator Effects



C. Mediator Effects

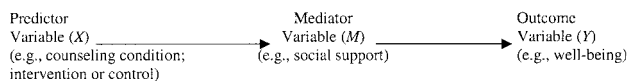


Figure 1. Diagrams of direct, moderator, and mediator effects.

way to do this is by examining moderators and mediators of these effects.

Questions involving moderators address “when” or “for whom” a variable most strongly predicts or causes an outcome variable. More specifically, a moderator is a variable that alters the direction or strength of the relation between a predictor and an outcome (Baron & Kenny, 1986; Holmbeck, 1997; James & Brett, 1984). Thus, a moderator effect is nothing more than an interaction whereby the effect of one variable depends on the level of another. For example, counseling researchers have long been admonished to investigate not only the general effectiveness of interventions but which interventions work best for which people (see Norcross, 2001, for a recent review of interaction effects in treatment outcome studies). For example, in Figure 1B, gender (variable *Z*) is introduced as a moderator of the relation between counseling condition and well-being. If gender is a significant moderator in this case, the counseling intervention increases well-being more for one gender than for the other (e.g., it increases well-being more for women than for men). Such interaction effects (i.e., moderators) are important to study because they are common in psychological research, perhaps even the rule rather than the exception (Jaccard et al., 1990). If moderators are ignored in treatment studies, participants may be given a treatment that is inappropriate or perhaps even harmful for them (Kraemer, Stice, Kazdin, Offord, & Kupfer, 2001).

Interaction effects are not only important for intervention studies, however. There are many other instances in which researchers are interested in whether relations between predictor and outcome variables are stronger for some people than for others. The identification of important moderators of relations between predictors and outcomes indicates the maturity and sophistication of a field of inquiry (Aguinis, Boik, & Pierce, 2001; Judd, McClelland, & Culhane, 1995) and is at the heart of theory in social science (Cohen et al., 2003). A recent example in *JCP* illustrates the ways in which examining moderator effects can increase our under-

standing of the relations between important predictors and outcomes. Specifically, Corning (2002) found that perceived discrimination was positively related to psychological distress only among individuals with low self-esteem (and not among individuals with high self-esteem). Thus, self-esteem “buffered” the effects of discrimination on distress.

Whereas moderators address “when” or “for whom” a predictor is more strongly related to an outcome, mediators establish “how” or “why” one variable predicts or causes an outcome variable. More specifically, a *mediator* is defined as a variable that explains the relation between a predictor and an outcome (Baron & Kenny, 1986; Holmbeck, 1997; James & Brett, 1984). In other words, a mediator is the mechanism through which a predictor influences an outcome variable (Baron & Kenny, 1986). In Figure 1C, social support (variable *M*) is introduced as a mediator of the relation between counseling condition and well-being. If social support is a significant mediator in this case, the reason the treatment group has higher well-being scores is that participants in this group report greater increases in social support than do those in the control condition. Alternatively, social support might be a significant mediator if those in the control condition reported greater decreases in social support than those in the treatment condition (i.e., the reason the treatment was effective was that it prevented decreases in social support). Within the context of evaluating counseling interventions, measuring underlying change mechanisms (i.e., mediators) as well as outcomes provides information on which mechanisms are critical for influencing outcomes (MacKinnon & Dwyer, 1993). This information can enable us to focus on the effective components of treatments and remove the ineffective components (MacKinnon, 2000) as well as to build and test theory regarding the causal mechanisms responsible for change (Judd & Kenny, 1981).

Furthermore, when there are chains of mediators, addressing only one link in the chain may limit treatment effectiveness, whereas sequential interventions that address each link may be more successful (Kraemer et al., 2001). As was the case with moderator research, testing mediators also is important outside of evaluating interventions. It is a sign of a maturing discipline when, after direct relations have been demonstrated, we have turned to explanation and theory testing regarding those relations (Hoyle & Kenny, 1999). For example, in a recent *JCP* article, Lee, Draper, and Lee (2001) found that the negative relation between social connectedness and distress was mediated by dysfunctional interpersonal behaviors. In other words, individuals low in connectedness reported more distress in part because they also engaged in more dysfunctional behaviors.

A given variable may function as either a moderator or a mediator, depending on the theory being tested. For example, social support could be conceptualized as a moderator of the relation between counseling condition and well-being. This would be the case if theory suggested that the intervention might be differentially effective for individuals high and low in social support. Social support also could be conceptualized as a mediator of the relation between counseling condition and well-being, as it is depicted in Figure 1C. In this case, theory would suggest that the reason counseling is effective is that it increases social support. Thus, the same variable could be cast as a moderator or a mediator, depending on the research question and the theory being tested. Although this can be confusing, it is helpful to keep in mind that

moderators often are introduced when there are unexpectedly weak or inconsistent relations between a predictor and an outcome across studies (Baron & Kenny, 1986). Thus, one might look for moderators if the evidence for the effectiveness of a given intervention is weak, which may be because it is effective only for some people. The choice of moderators should be based on a specific theory regarding why the intervention may be more effective for some people than for others. In contrast, one typically looks for mediators if there already is a strong relation between a predictor and an outcome and one wishes to explore the mechanisms behind that relation. In the counseling example, if there is solid evidence that an intervention is effective, one might want to test a specific theory about what makes the intervention effective. In short, decisions about potential moderators and mediators should be based on previous research and theory and are best made a priori in the design stage rather than post hoc.

One also can examine mediators and moderators within the same model. *Moderated mediation* refers to instances in which the mediated relation varies across levels of a moderator. *Mediated moderation* refers to instances in which a mediator variable explains the relation between an interaction term in a moderator model and an outcome. These more complex models have been described in more detail elsewhere (e.g., Baron & Kenny, 1986; Hoyle & Robinson, in press; James & Brett, 1984; Wegener & Fabrigar, 2000).

MODERATOR EFFECTS

Researchers can use multiple regression to examine moderator effects whether the predictor or moderator variables are categorical (e.g., sex or race) or continuous (e.g., age).² When both the predictor and moderator variables are categorical, analysis of variance (ANOVA) procedures also can be used, although multiple regression is preferred because of the flexibility in options it provides for coding categorical variables (Cohen et al., 2003). When one or both variables are measured on a continuous scale, regression procedures that retain the continuous nature of the variables clearly are preferred over using cut points (e.g., median splits) to create artificial groups to compare correlations between groups or examine interaction effects using ANOVA (Aiken & West, 1991; Cohen, 1983; Cohen et al., 2003; Jaccard et al., 1990; Judd et al., 1995; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993; West, Aiken, & Krull, 1996). This is because the use of cut points to create artificial groups from variables actually measured on a continuous scale results in a loss of information and a reduction in power to detect interaction effects.

However, artificially dichotomizing two continuous variables (e.g., a predictor and a moderator) also can have the opposite effect and can lead to spurious main and interaction effects (MacCallum et al., 2002). Simulation studies have shown that hierarchical multiple regression procedures that retain the true nature of continuous variables result in fewer Type I and Type II errors for detecting moderator effects relative to procedures that involve the use of cut points (Bissonnette, Ickes, Bernstein, & Knowles, 1990; Mason, Tu, & Cauce, 1996; Stone-Romero & Anderson, 1994). Statisticians also generally have encouraged the use of hierarchical regression techniques over the practice of comparing correlations between groups when the group variable is naturally categorical

(e.g., sex or race), because different correlations between groups may reflect differential variances between groups rather than true moderator effects (Baron & Kenny, 1986; Chaplin, 1991; Judd et al., 1995). Unfortunately, in contrast to these recommendations, MacCallum et al. (2002) concluded that *JCP* was one of three leading journals in psychology in which the dichotomization of continuous variables was a relatively common practice. Our review of research published in *JCP* in 2001 also suggested that the majority of researchers who tested interactions involving continuous variables dichotomized those variables and used ANOVA rather than regression.

Guide to Testing Moderator Effects in Multiple Regression

In this section, we first present a guide for using hierarchical multiple regression to examine moderator effects, including issues related to designing the study, analyzing the data, and interpreting the results. This is followed by a discussion of additional issues to consider when examining moderator effects using regression techniques. We then provide an example that illustrates the steps involved in performing a moderator analysis. Although we focus on testing moderator effects, many of the issues we raise apply to regression analyses more generally.

To identify aspects of testing moderation about which there may be confusion in counseling psychology research, we also performed a manual review of all articles published in the 2001 issues of *JCP*. In total, 54 articles appeared in 2001, including regular articles, comments, replies, and brief reports. Of these 54, 12 (22%) contained a test of an interaction (although the study was not always framed as a test of a moderation hypothesis). Only 4 of the 12 used multiple regression with an interaction term to test moderation, which is the procedure we describe subsequently. Although this sample of articles is small, our reading of these articles suggested points of confusion, as noted in the discussion to follow. The results of our review are reported on a general level to avoid singling out particular studies or authors.

Designing a Study to Test Moderation

Importance of Theory

All of the study design decisions outlined next should be made on the basis of a well-defined theory, which unfortunately is not often the case (Chaplin, 1991). For example, both the choice of a moderator and the hypothesized nature of the interaction should be based on theory (Jaccard et al., 1990). Cohen et al. (2003, pp. 285–286) described three patterns of interactions among two continuous variables: enhancing interactions (in which both the predictor and moderator affect the outcome variable in the same direction and together have a stronger than additive effect), buffering interactions (in which the moderator variable weakens the

² Baron and Kenny (1986) distinguished situations in which the predictor is continuous and the moderator is categorical from situations in which the predictor is categorical and the moderator is continuous. However, the analyses are the same if it is assumed, as typically is the case, that the effect of the predictor on the outcome variable changes linearly with respect to the moderator.

effect of the predictor variable on the outcome), and antagonistic interactions (in which the predictor and moderator have the same effect on the outcome but the interaction is in the opposite direction). Similarly, in the case of one categorical and one continuous variable, the theory on which the hypotheses are based may specify that a predictor is positively related to an outcome for one group and unrelated for another group. Alternatively, theory may specify that a predictor is positively related to outcomes for one group and negatively related to outcomes for another. Finally, the interaction may be nonlinear and thus not captured by a simple product term.³ In the *JCP* articles we reviewed, the specific nature of the interaction rarely was specified a priori.

Power of Tests of Interactions

Although hierarchical multiple regression appears to be the preferred statistical method for examining moderator effects when either the predictor or the moderator variable (or both) is measured on a continuous scale (Aguinis, 1995), concerns often have been raised in the statistical literature about the low power of this method to detect true interaction effects. Aguinis et al. (2001) showed that the power to detect interaction effects in a typical study is .20 to .34, much lower than the recommended level of .80. Low power is a particular problem in nonexperimental studies, which have much less power for detecting interaction effects than do experiments (McClelland & Judd, 1993). All but one of the studies that we reviewed in *JCP* was nonexperimental, and none of the authors reported the power of the test of the interaction.

Several factors have been identified that reduce the power of tests of interactions. These factors, outlined next, should be taken into consideration when designing a study to test moderator effects to increase the chances of finding significant interactions when they exist. Otherwise, it is unclear whether the interaction is not significant because the theory was wrong or the test of the interaction lacked sufficient power. As discussed by Aguinis (1995), the importance of fully considering issues related to research design and measurement before data are collected cannot be overstated.

Effect size for interaction and overall effect size. To ensure adequate sample sizes to maximize the chances of detecting significant interaction effects, the size of the interaction effect should be estimated before data collection. To be specific, the effect size for the interaction in a regression analysis is the amount of incremental variance explained by the interaction term after the first-order effects have been controlled (i.e., the R^2 change associated with the step in which the interaction term is added). Thus, the pertinent research should be reviewed so that the expected effect size can be estimated on the basis of what is typically found in the literature. Generally, effect sizes for interactions are small (Chaplin, 1991), as was the case for the studies we reviewed in *JCP*. According to Cohen's (1992) conventions, a small effect size in multiple regression corresponds to an R^2 value of .02. The sample size needed to detect a moderator effect depends on the size of the effect; if the interaction effect is small, a relatively large sample is needed for the effect to be significant. Methods for calculating needed sample sizes are discussed in a later section, because the sample size needed for adequate power depends on several factors other than the effect size of the interaction.

In addition to the size of the interaction effect, the total effect size (i.e., the amount of variance explained by the predictor, moderator, and interaction) should be estimated before data collection. Again, this is done by reviewing the pertinent literature to determine how much variance typically is accounted for by the variables included in one's model. Neither the interaction effect size nor the total effect size was estimated a priori in any of the studies we reviewed in *JCP*. Moderator effects are best detected (i.e., tests have more power) when the relation between the predictor and outcome is substantial (Chaplin, 1991; Jaccard et al., 1990). However, moderators often are examined when there are unexpectedly weak relations between a predictor and outcome (Baron & Kenny, 1986; Chaplin, 1991), which further contributes to the low power of many tests of interactions. One suggested way to increase power is to increase the multiple correlation between the full model and the outcome variable by including additional significant predictors of the outcome variable in the model as covariates (Jaccard & Wan, 1995).

Choosing variables. Keeping in mind that decisions regarding tests of interactions should be based on theory, there are several factors to consider with regard to choosing predictor, moderator, and outcome variables, each of which can increase or decrease the power of interaction tests. Somewhat different issues arise with regard to categorical variables, continuous variables, and outcome variables. Issues associated with each type of variable are discussed in turn.

There are two issues to consider with regard to categorical variables. The first is that unequal sample sizes across groups decrease power (Aguinis, 1995; Aguinis & Stone-Romero, 1997; Alexander & DeShon, 1994; Stone-Romero, Alliger, & Aguinis, 1994). For example, with two groups, power decreases as the sample size proportions vary from .50/.50, regardless of the total sample size. With a sample size of 180, the power to detect a difference of .40 in a correlation between two groups (e.g., a correlation between a predictor and outcome of .2 for women and .6 for men) is more than .80 if the two groups are equal in size. However, if the sample size proportion is .10/.90 (e.g., 10% men and 90% women), power is about .40 (Stone-Romero et al., 1994). If the categorical variable is an experimental condition (e.g., type of counseling intervention), this can be addressed by assigning equal numbers of individuals to each group. However, when the categorical variable is not manipulated (e.g., gender or race), unequal groups are likely, and the effects on power need to be evaluated. Indeed, in some of the *JCP* studies reviewed, proportions were as skewed as .07/.93, although this inequality was never mentioned as a potential problem with regard to power.

A second issue to consider is that even if sample sizes are equal across groups, error variances across groups may be unequal (DeShon & Alexander, 1996; Overton, 2001). In fact, one review revealed that the assumption of homogeneous error variance is violated about half of the time (Aguinis, Petersen, & Pierce, 1999). If sample sizes and error variances are unequal, power can be

³ In this article, we focus on linear interactions because they are the most common form of interaction tested (for information on nonlinear interactions, see Aiken & West, 1991, chap. 5; Cohen et al., 2003, chaps. 7 and 9; Jaccard et al., 1990, chap. 4; Lubinski & Humphreys, 1990; MacCallum & Mar, 1995).

either overestimated or underestimated, depending on whether the larger or smaller sample has the larger error variance (for more details, see Aguinis & Pierce, 1998; Grissom, 2000; Overton, 2001). In these cases, the results of multiple regression analyses cannot be trusted, and alternative tests should be used. Aguinis et al. (1999) developed a program, available on the World Wide Web, that both tests the assumption of homogeneous error variance and calculates alternative tests.⁴ These alternative tests make a practical difference when the error variance of one group is 1.5 times larger than that of the other group (DeShon & Alexander, 1996; Overton, 2001). Only one study in our *JCP* review reported whether the assumption of homogeneity of error variance had been met.

There also are two issues to consider when choosing continuous variables. One is the reliability of the measures. Measurement error in individual variables (either predictors or moderators) dramatically reduces the reliability of the interaction term constructed from them (Aguinis, 1995; Aguinis et al., 2001; Aiken & West, 1991; Busemeyer & Jones, 1983; Jaccard et al., 1990). Lower reliability of the interaction term increases its standard error and reduces the power of the test. For example, Aiken and West showed that the power of the test of the interaction is reduced by up to half with reliabilities of .80 rather than 1.00. The second issue concerns restriction in range, which also reduces power (Aguinis, 1995; Aguinis & Stone-Romero, 1997; McClelland & Judd, 1993). Range restriction means that not all individuals in a population have an equal probability of being selected for the sample (Aguinis, 1995). A simulation study examining the effects of several variables on power showed that range restriction had a considerable effect (Aguinis & Stone-Romero, 1997). McClelland and Judd (1993) provided specific recommendations regarding oversampling techniques that can be used to address this issue (see also Cohen et al., 2003, pp. 298–299). In the 2001 *JCP* studies we reviewed, most measures had adequate reliability (.80 or higher), although range restriction was rarely mentioned as a possible issue and was difficult to assess because adequate information (e.g., means, standard deviations, skewness, and ranges) was not always provided.

A final consideration is choice of an outcome variable. Lower reliability of an outcome variable reduces correlations with predictors, thus lowering the overall R^2 value and the power of the test (Aguinis, 1995). Furthermore, if the outcome measure does not have enough response options (i.e., is too “coarse”) to reflect the interaction, there will be a loss in power (Russell & Bobko, 1992). The outcome measure has to have as many response options as the product of the response options of the predictor and moderator variables. For example, if both the predictor and moderator are measured with 5-point Likert scales, the true moderator effect will contain 5×5 conceptually distinct latent responses. The outcome measure will thus need to have 25 response options (25-point scale) to capture the true moderator effect. Russell and Bobko also noted that summing responses to multiple Likert-type items with limited response options (e.g., 5-point scale) does not provide the same increase in power as using an outcome measure with more response options (e.g., 25-point scale) because participants are still responding to each item on the limited scale. In other words, the number of response options for the items determines coarseness rather than the number of items on the scale. Because many outcome measures do not have sufficient response options, the

effects of scale coarseness on power may be difficult to avoid. Aguinis, Bommer, and Pierce (1996) developed a computer program that administers questionnaires by prompting respondents to click along a line on a computer screen, thus allowing for more response options and increasing the accuracy of tests of moderator effects. However, if researchers prefer to use measures with established reliability and validity, they need to recognize that scale coarseness may decrease power to detect the interaction. Scale coarseness was never mentioned as a factor that may affect power in the *JCP* studies we reviewed.

There are several resources to which researchers can turn to estimate power. Jaccard et al. (1990) and Aiken and West (1991) provided tables for estimating power for interactions. There is also an online calculator that estimates the sample size needed to achieve a given level of power with categorical moderators that takes into account many of the factors just listed (e.g., sample size of each group, effect size of interaction, and reliability of measures).⁵ This program can be used a priori to assess the effect of various design decisions (e.g., to maximize power, is it better to sacrifice reliability or sample size? see Aguinis et al., 2001, for further details).

In summary, to maximize the power of tests of moderator effects, researchers are encouraged to rely on theory when planning moderator analyses, use an experimental design when appropriate, determine and obtain the sample size needed to achieve adequate power based on estimated effect sizes and other factors, attempt to collect equal numbers of participants for different levels of a categorical variable, test the homogeneity of error variance assumption and use appropriate tests if it is violated, choose highly reliable continuous variables, obtain measures of continuous predictor and moderator variables that are normally distributed, and use outcome measures that are both reliable and sufficiently sensitive (i.e., have enough scale points to capture the interaction). Some (Aguinis, 1995; Jaccard & Wan, 1995; Judd et al., 1995; McClelland & Judd, 1993) also have suggested raising the alpha level above the traditional .05 level to maximize power, with various caveats.

Although these practices would improve the probability that researchers would find significant moderator effects when they exist, they may not always be possible to implement. In addition, there may be times when other statistical procedures may be more appropriate because of limitations inherent in ordinary least squares regression. Most notably, several authors (e.g., Aguinis, 1995; Aiken & West, 1991; Baron & Kenny, 1986; Busemeyer & Jones, 1983; Holmbeck, 1997; Jaccard et al., 1990) have encouraged the use of SEM as a way to control for unreliability in measurement. SEM can be used to examine interactions involving both categorical and continuous variables (for details on how to perform such analyses, see Bollen & Paxton, 1998; Holmbeck, 1997; Jaccard & Wan, 1995, 1996; Kenny & Judd, 1984; Moulder & Algina, 2002; Ping, 1996; Schumacker & Marcoulides, 1998). When one variable is categorical, a multiple-group approach can be used in which the relation between the predictor and outcome is

⁴ The program can be found at <http://members.aol.com/imsap/altmmr.html>.

⁵ This program can be found at <http://www.math.montana.edu/~rjboik/power.html>.

estimated separately for the multiple groups. Specifically, an unconstrained model is compared with a constrained model (in which the paths are constrained to be equal across groups). If the unconstrained model is a better fit to the data, there is evidence of moderation (i.e., different relations between the predictor and outcome across groups). However, SEM techniques for testing interactions between continuous variables are complex, and there is little consensus regarding which of several approaches is best (Marsh, 2002).

Analyzing the Data

After the study has been designed and the data collected, the data need to be analyzed. Steps involved in analyzing the data include creating or transforming predictor and moderator variables (e.g., coding categorical variables, centering or standardizing continuous variables, or both), creating product terms, and structuring the equation.

Representing Categorical Variables With Code Variables

If either the predictor or moderator variable is categorical, the first step is to represent this variable with code variables. The number of code variables needed depends on the number of levels of the categorical variable, equaling the number of levels of the variable minus one. For example, a counseling outcome study in which participants are randomly assigned to one of three treatment conditions (e.g., cognitive-behavioral therapy, interpersonal therapy, and control group) would need two code variables to fully represent the categorical variable of treatment type in the regression equation. One of several coding systems can be chosen to represent the categorical variable based on the specific questions being examined (West et al., 1996). Specifically, dummy coding is used when comparisons with a control or base group are desired, effects coding is used when comparisons with the grand mean are desired, and contrast coding is used when comparisons between specific groups are desired.

Using the three-condition treatment study as an example, dummy coding would be used to compare the mean of each therapy group with the mean of the control group, effects coding would be used to compare each of the group's means with the grand mean, and contrast coding would be used to compare orthogonal combinations of the categorical variable (e.g., comparisons of the mean of the two treatment groups with the mean of the control group and comparisons of the means of each treatment group with each other). We discuss this in more detail in our example, but it is critical to note here that the choice of coding system has very important implications for testing and interpreting effects in equations involving interactions. We refer readers to West et al. (1996), in particular, for a complete discussion of the differences among coding systems and practical guidelines regarding when and how to use them (see also Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990). In the *JCP* articles we reviewed, only dummy coding was used. However, as noted by Cohen et al. (2003), "the dummy coding option that is so often considered the 'default' will frequently not be the optimal coding scheme" (p. 375).

Centering or Standardizing Continuous Variables

The next step in formulating the regression equation involves centering or standardizing predictor and moderator variables that are measured on a continuous scale.⁶ Several statisticians recommend that these variables be centered (i.e., put into deviation units by subtracting their sample means to produce revised sample means of zero). This is because predictor and moderator variables generally are highly correlated with the interaction terms created from them. Centering reduces problems associated with multicollinearity (i.e., high correlations) among the variables in the regression equation (for further explanation, see Cohen et al., 2003; Cronbach, 1987; Jaccard et al., 1990; West et al., 1996). There may be further benefits to standardizing (i.e., *z* scoring) rather than centering continuous predictor and moderator variables (Aiken & West, 1991; Friedrich, 1982). For example, standardizing these variables makes it easier to plot significant moderator effects because convenient representative values (i.e., the mean and ± 1 standard deviation from the mean) can be substituted easily into a regression equation to obtain predicted values for representative groups when the standard deviations of these variables equal one (see Cohen et al., 2003). In addition, *z* scores are very easy to create within standard statistical packages. Standardizing also makes it easier to interpret the effects of the predictor and moderator, as we discuss later. In contrast to these recommendations, only one of the *JCP* articles reviewed reported using centered or standardized continuous variables.

Creating Product Terms

After code variables have been created to represent any categorical variables and variables measured on a continuous scale have been centered or standardized, product terms need to be created that represent the interaction between the predictor and moderator. To form product terms, one simply multiplies together the predictor and moderator variables using the newly coded categorical variables or centered/standardized continuous variables (Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990; West et al., 1996). A product term needs to be created for each coded variable (e.g., if there is one coded variable for a categorical variable with two levels, there is one interaction term; if there are two coded variables for a categorical variable with three levels, there are two interaction terms). This product term does not need to be centered or standardized.

Structuring the Equation

After product terms have been created, everything should be in place to structure a hierarchical multiple regression equation using standard statistical software to test for moderator effects. To do this, one enters variables into the regression equation through a series of specified blocks or steps (Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990; West et al., 1996). The first step

⁶ Whereas there are benefits to centering or standardizing predictor variables and moderator variables that are measured on a continuous scale, there typically is no reason to do so with code variables representing categorical variables or continuous outcome variables (Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990; West et al., 1996).

generally includes the code variables and centered/standardized variables representing the predictor and the moderator variables. All individual variables contained in the interaction term(s) must be included in the model (West et al., 1996). Product terms must be entered into the regression equation after the predictor and moderator variables from which they were created (Aiken & West, 1991; Cohen et al., 2003; Dunlap & Kemery, 1987; Holmbeck, 1997; Jaccard et al., 1990; McClelland & Judd, 1993; West et al., 1996). Inspecting product terms by themselves (without controlling for the variables from which they are based) confounds the moderator effect with the effects of the predictor and moderator variables (Judd et al., 1995). If two or more product terms have been created because a categorical variable has more than two levels, all of the product terms should be included in the same step (Aiken & West, 1991; Jaccard et al., 1990; West et al., 1996).

Interpreting the Results

Interpreting the results of hierarchical multiple regression analyses that examine a moderator effect involves the following: (a) interpreting the effects of the predictor and moderator variables, (b) testing the significance of the moderator effect, and (c) plotting significant moderator effects.

Interpreting the Effects of the Predictor and Moderator Variables

The interpretation of regression coefficients representing the relations between the predictor and the outcome variable and between the moderator and the outcome variable is unique in multiple regression models examining moderator effects. That is, such relations are interpreted as “conditional” effects at the value of 0 for the other variables included in the model and not as “main effects,” as is often the practice in published studies (Judd et al., 1995). For example, if social support is the predictor, gender is the moderator, and depression is the outcome, the regression coefficient for social support represents the regression of depression on social support when gender is coded 0. Likewise, the regression coefficient for gender represents the regression of depression on gender when social support is 0. If, for example, social support is measured on a Likert scale with response options of 1–5, the regression coefficient for gender would represent the regression of depression on gender at a value not defined for social support (i.e., 0). This is another reason why predictor and moderator variables should be centered or standardized; doing so provides a meaningful zero point (i.e., the mean for continuous variables) for these interpretations (see Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990; Judd et al., 1995; and West et al., 1996, for further discussion).⁷ In other words, when variables are centered or standardized, the first-order effect of one variable represents the effect of that variable at the average level of the other variable(s). That first-order effects are conditional was never mentioned in the *JCP* studies we reviewed.

Another point to mention regarding the interpretation of first-order effects in equations containing interactions occurred to us after reviewing the *JCP* articles. On a few occasions, researchers interpreted the first-order effects before the interaction term had been entered into the equation. Whether this is appropriate depends on the underlying theoretical model (Cohen et al., 2003). Consider,

for example, a case in which a predictor variable is entered in the first step, a moderator variable is entered in the second step, and the multiplicative term reflecting the interaction between them is entered in the third step. If the regression coefficient for the predictor variable in the first step is interpreted, all of the variance shared among the predictor, the moderator, and their interaction is attributed to the predictor. This is justified only if a strong theoretical argument can be made that the predictor causes the moderator (see Cohen et al., 2003, for approaches to testing the significance of different elements in a regression equation containing interactions).

Finally, when interpreting the results, it is important to note that one should interpret the unstandardized (B) rather than standardized (β) regression coefficients because, in equations that include interaction terms, the β coefficients for the interaction terms are not properly standardized and thus are not interpretable (see Aiken & West, 1991, pp. 40–47, for a detailed rationale; see also Cohen et al., 2003; West et al., 1996). Whether this was done correctly in the *JCP* articles we reviewed was unclear, because specific coefficients sometimes were not reported and because of confusion in terminology (e.g., not following the convention of referring to unstandardized coefficients as B s and standardized coefficients as β s [betas]).

Testing the Significance of the Moderator Effect

The method of determining the statistical significance of the moderator effect depends to some degree on the characteristics of the predictor and moderator variables. That is, when a moderator effect is composed of predictor and moderator variables that both are measured on a continuous scale, one continuous variable and one categorical variable with two levels, or two categorical variables each with two levels, the single degree of freedom F test, representing stepwise change in variance explained as a result of the addition of the product term, provides the information needed (Aiken & West, 1991; Jaccard et al., 1990; West et al., 1996). This process is somewhat different, however, when categorical variables have more than two levels. As discussed most clearly by West et al. (1996; see also Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990), at least two code variables are needed to fully represent the categorical variable in this situation. Consequently, the moderator effect is tested with the multiple degree of freedom omnibus F test representing stepwise change for the step in which the multiple product terms are entered. If the omnibus F test is statistically significant, the single degree of freedom t tests related to specific product terms are inspected to determine the form of the moderator effect. The importance of specific comparisons also can be conceptualized in terms of the amount of variance accounted for (i.e., by their squared semipartial correlations; see Cohen et al., 2003). In other words, when there is more than one coded variable, the amount of variance in the outcome variable accounted for by each comparison is indexed by the squared semipartial correlation associated with that comparison. Cohen et

⁷ Similarly, dummy coding using values other than 0 and 1 (i.e., 1 and 2) is not recommended in regression models involving interactions, because 0 is not a valid value when a 1, 2 coding scheme is used (West et al., 1996).

al. described how to calculate semipartial correlations, which are not always provided by statistical programs in their standard output, and provided tables to calculate the power associated with tests of their significance.

If the interaction term is not significant, the researcher must decide whether to remove the term from the model so that the first-order effects are not conditional effects. Aiken and West (1991, pp. 103–105) reviewed the issues associated with this decision and ultimately recommended keeping the nonsignificant interaction term in the model if there are strong theoretical reasons for expecting an interaction and removing the interaction if there is not a strong theoretical rationale for the moderator effect (see also Cohen et al., 2003).

Interpreting Significant Moderator Effects

Once it has been determined that a significant moderator effect exists, it is important to inspect its particular form. There are two ways to do this. The first is to compute predicted values of the outcome variable for representative groups, such as those who score at the mean and 1 standard deviation above and below the mean on the predictor and moderator variables (Aiken & West, 1991; Cohen et al., 2003; Holmbeck, 1997; West et al., 1996). The predicted values obtained from this process then may be used to create a figure summarizing the form of the moderator effect. The second method is to test the statistical significance of the slopes of the simple regression lines representing relations between the predictor and the outcome at specific values of the moderator variable (for further details, see Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990; West et al., 1996). Unlike just plotting means, testing the simple slopes provides information regarding the significance of the relations between the predictor and outcome at different levels of the moderator. Confidence intervals for the simple slopes also can be calculated (Cohen et al., 2003). Among the *JCP* articles we reviewed, only one provided plots of interactions (which were mislabeled), and none presented tests of simple slopes.

Additional Issues to Consider When Examining Moderator Effects

Having discussed the basics of investigating moderator effects using hierarchical multiple regression techniques, we now turn to some additional issues that may be important to consider: (a) including covariates in regression equations examining moderator effects, (b) examining multiple moderator effects, and (c) examining three-way (and higher) interactions.

Including Covariates in Regression Equations Examining Moderator Effects

In addition to variables needed to test a moderator effect, some researchers may want to consider including covariates to control for the effects of other variables, increase the overall R^2 to increase power, or estimate change in an outcome variable over time. If this is to be done, covariates need to be entered in the first step of the regression equation, followed by the predictor variable, moderator variable, and product terms in subsequent steps, as discussed earlier. In addition, as emphasized by Cohen and Cohen (1983),

interactions between covariates and other variables in the regression model should be tested to determine whether covariates act consistently across levels of the other variables (i.e., have parallel slopes).⁸ This may be done by adding a final step containing interactions between the covariates and all other variables (including product terms). If the omnibus F test representing this entire step is not significant, this step can be dropped from the model. If the overall step is significant, the t tests related to specific interactions can be inspected, potentially uncovering a moderator effect that can be investigated in future research (Aiken & West, 1991; Cohen & Cohen, 1983). In the *JCP* articles we reviewed, when covariates were added to regression models containing interactions, interactions with covariates were never assessed.

Examining Multiple Moderator Effects

Although we have been focusing on models with one moderator variable, some researchers may want to consider investigating multiple moderator effects. However, performing a large number of statistical tests in this manner will lead to an inflated Type I error rate (Cohen et al., 2003). To help control for this type of error, all of the moderator effects being considered may be entered in a single step after all of the predictor and moderator variables from which they are based have been entered in previous steps. The significance of the omnibus F test representing the variance explained by this entire step then can determine whether it should be eliminated from the model (if the omnibus test is not significant) or whether t tests representing specific moderator effects should be inspected for statistical significance (if the omnibus test is significant; Aiken & West, 1991). The squared semipartial correlations associated with each interaction also can be calculated to determine the amount of variance in the outcome attributable to each interaction term (Cohen et al., 2003). Significant moderator effects then can be explored in the manner discussed earlier. When multiple moderators were tested in the *JCP* articles we reviewed, inflated Type I error was never mentioned or addressed.

Examining Higher Order Interactions

We focus in this article on procedures for testing two-way interactions because they are the most common form of interaction hypothesized and tested in counseling psychology research. Indeed, higher order interactions were never tested in the *JCP* articles we reviewed. However, interactions may involve three (or more) variables. To use the example we work through later, the relation between social support and depression may depend on age as well as gender. Researchers interested in testing higher order interactions are referred to Aiken and West (1991, chap. 4), who provided a thorough discussion of the procedures for testing and interpreting three-way interactions (see also Cohen et al., 2003). Some caveats regarding higher order interactions should be mentioned, however. For example, three-way (and higher order) interactions are rarely of primary interest because our theories are not sufficiently complex (Cohen & Cohen, 1983). As mentioned be-

⁸ Because the covariates will be entered into interaction terms, it is useful to center or standardize them. Even if they are not entered into interaction terms, Cohen et al. (2003) recommended centering them to be consistent with other predictors in the model.

fore, all tests of moderation should be based on strong theory, with the nature of the interaction specified a priori. No interactions should be included unless they have substantive theoretical support because as the number of hypotheses tested increases, so do the risks of Type I and Type II error (Chaplin, 1991; Cohen & Cohen, 1983; McClelland & Judd, 1993). Also, measurement error is an even bigger problem for three-way than for two-way interactions (Busemeyer & Jones, 1983) because the reliability of the product term is the product of the reliability of the three measures.

Conclusions

There are many issues to consider in testing moderation and many issues about which counseling psychology researchers seem to be unaware. Indeed, few researchers in the studies we reviewed tested moderation, and those who did used methods other than multiple regression. For example, it was common for researchers to dichotomize continuous variables and use other analytic approaches (such as ANOVA), resulting in a loss of power and information. Among those who did use regression to test moderation, little or no effort was made to estimate power or to address issues that may lower power (e.g., unequal sample sizes). Furthermore, there appeared to be little awareness of issues involved in interpreting results from moderational analyses, such as the need to center continuous variables.

Example: Testing Moderator Effects Using Multiple Regression

To illustrate how moderator effects may be investigated through the use of multiple regression, we provide a step-by-step example using simulated data that meet the criteria outlined previously. In the mediation example provided later, actual data are used to illustrate some of the issues that arise when using real, versus simulated, data.

Designing the Study

Recall that it often is useful to look for moderators when there are unexpectedly weak or inconsistent relations between a predictor and an outcome across studies. One example of this is the relations between social support and mental health indicators (e.g., depression), which often are not as strong as one might expect (e.g., see Lakey & Drew, 1997). Thus, perhaps it is the case that social support is more strongly related to depression for some people than for others. On the basis of existing theory and research, one possible moderator of the relation between social support and depression is gender. Specifically, because relationships generally are more important to women than to men (Cross & Madson, 1997), the relation between social support and depression may be stronger for women than for men. In our example, we measured social support in terms of unhelpful social support (e.g., minimizing the event), which tends to be more strongly related to depression than is helpful support (e.g., Frazier, Tix, & Barnett, 2003). Thus, the hypothesis tested in this example, based on previous research and theory, is that unhelpful support behaviors will be positively related to depression for both men and women but that this relation will be stronger for women than for men.

We also took into account the factors mentioned earlier that affect power and incorporated several design features to maximize

power. For example, we generated the data such that the range in the social support measure (the predictor) was not restricted and the reliability coefficients for the social support and depression (the outcome) measures were good (i.e., alpha coefficients of .80). The social support measure contained 20 items rated on a 5-point Likert-type scale. We verified that the homogeneity of error variance assumption was not violated using an online calculator (see Aguinis et al., 1999, and Footnote 4). With regard to effect size, we generated the data such that there would be a difference of about .30 in the correlation between unhelpful support and depression for women ($r = .39$) and men ($r = .10$). To estimate the sample size needed to have sufficient power (.80) to detect this effect, we entered these parameters into an online power calculator (see Aguinis et al., 2001, and Footnote 5) and determined that we would need a sample of about 160 in each group. The data set generated consisted of equal numbers of men and women (165 in each group), with an actual power of .83 to detect the specified differences. Our outcome measure consisted of 20 items rated on a 10-point Likert scale. Because there were 10 response options for each item on the outcome measure, it was sensitive enough (i.e., not too coarse) to capture the interaction between social support (5 response options) and gender (2 response options). (Recall that the number of response options for the outcome variable should be greater than or equal to the product of the number of response options for the predictor and moderator variables.) Finally, the one aspect that reduced our power was the use of a nonexperimental design. Although experimental designs have more power, our example involved a correlational study, because this is the design most often used to examine moderator effects in counseling psychology research.

Analyzing the Data

First, we standardized the unhelpful support variable so that it had a mean of 0 and a standard deviation of 1. Next, we needed to decide which form of coding to use for our categorical moderator variable (gender). We chose effects coding because we wanted to interpret the first-order effects of gender and social support as average effects, as in ANOVA (see Cohen et al., 2003, and West et al., 1996, for more details). More specifically, if one codes gender using effects coding (i.e., codes of -1 for men and 1 for women), and if the social support measure has been standardized so that it has a mean of 0, the first-order effect of gender is the average relation between gender and depression, the first-order effect of social support is the average relation between social support and depression, and the intercept is the average depression score in the sample. Because we had equal numbers of men and women in our sample, weighted and unweighted effects coding would give the same results. West et al. provided guidelines regarding when to use weighted versus unweighted effects coding if sample sizes are unequal. Because there were only two categories for gender, we needed only one code variable.⁹ The final step was to create the interaction term (the product of the gender code and the z -scored unhelpful support measure). To perform the

⁹ Readers are referred to the following sources for guidance on analyzing categorical data with more than two levels: Aiken and West (1991), Cohen et al. (2003), Jaccard et al. (1990), and West et al. (1996).

Table 1
Testing Moderator Effects Using Hierarchical Multiple Regression

Step and variable	<i>B</i>	<i>SE B</i>	95% CI	β	R^2
Effects coding (men coded -1, women coded 1)					
Step 1					
Gender	-0.11	0.07	-0.25, 0.02	-.09	
Unhelpful social support (<i>z</i> score)	0.32	0.07	0.18, 0.45	.25**	.07**
Step 2					
Gender \times Unhelpful Social Support	0.20	0.07	0.06, 0.33	.16*	.02*
Dummy coding (men coded 0, women coded 1)					
Step 1					
Gender	-0.23	0.13	-0.49, 0.03	-.09	
Unhelpful social support (<i>z</i> score)	0.12	0.09	-0.06, 0.30	.10	.07**
Step 2					
Gender \times Unhelpful Social Support	0.39	0.13	0.13, 0.65	.22*	.02*
Dummy coding (women coded 0, men coded 1)					
Step 1					
Gender	0.23	0.13	-0.03, 0.49	.09	
Unhelpful social support (<i>z</i> score)	0.51	0.10	0.32, 0.70	.40**	.07**
Step 2					
Gender \times Unhelpful Social Support	-0.39	0.13	-0.65, -0.13	-.22*	.02*

Note. CI = confidence interval.

* $p < .01$. ** $p < .001$.

analysis, we regressed depression on gender and the *z*-scored support measure in the first step and the interaction between gender and the *z*-scored support measure in the second step. The output is presented in Table 1.

Interpreting the Results

First, we obtained descriptive statistics to verify that the gender variable was coded correctly and that the social support variable had a mean of 0 and a standard deviation of 1.¹⁰ We also obtained correlations among all variables to make sure that, as a result of standardizing continuous variables, the interaction term and its components were not too highly correlated. As mentioned, multicollinearity can cause both interpretational and computational problems.

Looking at the output for effects coding in Table 1, the unstandardized regression coefficient for gender was -0.11, which was not significant at the conventional .05 level ($p = .09$). The unstandardized regression coefficient for unhelpful social support was 0.32 ($p < .0001$), meaning that there was a significant positive relation between unhelpful support and depression in the sample. Because gender was coded by means of effects coding, and the support variable was standardized, we could interpret this first-order effect of social support as an average effect. This would not be the case if another form of coding had been used. The unstandardized regression coefficient for the interaction term was .20 ($p = .004$). The R^2 change associated with the interaction term was .02. In other words, the interaction between unhelpful social support and gender explained an additional 2% of the variance in depression scores over and above the 7% explained by the first-order effects of social support and gender alone.

To understand the form of the interaction, it was necessary to explore it further. As mentioned, one way is to plot predicted values for the outcome variable (depression) for representative groups. A common practice (recommended by Cohen et al., 2003) is to choose groups at the mean and at low ($-1 SD$ from the mean) and high ($1 SD$ from the mean) values of the continuous variable. Here we plotted scores for men and women at the mean and at low ($-1 SD$) and high ($1 SD$) levels of unhelpful social support (see Figure 2). (If we had two continuous variables, we could plot scores for participants representing the four combinations of low and high scores on the two variables.) Predicted values were obtained for each group by multiplying the respective unstandardized regression coefficients for each variable by the appropriate value (e.g., -1, 1 for standardized variables) for each variable in the equation.¹¹ For example, to get the predicted score for men who score 1 standard deviation above the mean on unhelpful social support, we multiplied the unstandardized coefficient for gender (-0.11) by -1 (the code for men), multiplied the unstandardized coefficient for unhelpful support ($B = 0.32$) by 1 (the code for high levels of unhelpful social support), multiplied the unstandardized coefficient for the interaction term ($B = 0.20$) by the product of the gender and unhelpful support codes (in this case, $-1 \times 1 = -1$), and added the constant (5.10) for a predicted value on the depression measure of 5.34. The group with the lowest level of depres-

¹⁰ A scale may no longer be properly standardized if there are missing data and the sample from which the *z* score was created differs from the sample for the regression analyses.

¹¹ An Excel file created to calculate these predicted values is available from Patricia A. Frazier.

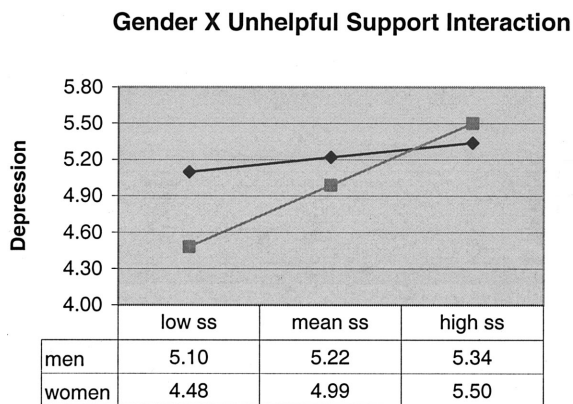


Figure 2. Plot of significant Gender × Unhelpful Social Support (ss) interaction. Solid diamonds = men; solid squares = women.

sion was women with low levels of unhelpful support ($Y = 4.48$), whose depression score was lower than that of men with low levels of unhelpful support ($Y = 5.10$). Men ($Y = 5.34$) and women ($Y = 5.50$) in the groups at high levels of unhelpful support had very similar depression scores, as did men ($Y = 5.22$) and women ($Y = 4.99$) with mean levels of unhelpful support.

Another approach is to test the significance of the slopes for each group. The significant interaction term tells us that the slopes differ from each other but not whether each slope differs from zero. For example, looking at Figure 2, we can formally test whether the slope representing the relation between social support and depression for men significantly differs from zero and whether the slope for women significantly differs from zero. To test the simple slopes for each group, we needed to conduct two additional regression analyses. Although these regressions were similar to those just reported, we recoded gender using dummy coding. In one analysis gender was coded so that men received a value of 0, and in one gender was coded so that women received a value of 0 (see Table 1).

As discussed earlier, when regression equations contain interaction terms, the regression coefficient for the predictor represents the relation between the predictor and outcome when the moderator has a value of 0. Thus, with gender dummy coded and men coded as 0, the regression coefficient for unhelpful social support represents the relation between unhelpful support and depression for men. With gender dummy coded and women coded as 0, the regression coefficient for unhelpful social support is the relation between unhelpful support and depression for women. When these two regressions were performed, there was a significant positive slope for women ($B = 0.51, p < .0001$) but not for men ($B = 0.12, p = .20$). As can be seen in Table 1, the 95% confidence interval for the simple slope for men included zero, which means that we could not reject the null hypothesis that this slope differed from zero. If we had not included gender as a moderator, we would have concluded that unhelpful social support had a small to medium-sized relation with depression ($B = .32$), which would have masked the fact that the relation was much stronger in women ($B = .51$) than in men ($B = .12$).¹² These analyses also illustrate how the coding of the gender variable changes the regression coefficients for the social support variable (but not the variance accounted for by the interaction term; see Cohen et al., 2003).

Now that we have found a significant interaction between gender and unhelpful support in predicting depression, what do we do? One possibility is to examine what accounts for the gender difference in the relation between unhelpful support and depression. That is, why is unhelpful support more related to depression for women than for men? Earlier we hypothesized that a possible reason is that relationships tend to be more important for women than for men (Cross & Madson, 1997). Thus, in a future study we could assess whether differences in the importance of relationships mediate the interaction between gender and social support in predicting depression. This would be an example of mediated moderation (Baron & Kenny, 1986). That gender moderates the relations between unhelpful support and depression also may have implications for interventions (i.e., interventions that improve social relationships may be more helpful for women than for men).

MEDIATOR EFFECTS

We now turn to a description of testing mediator effects, using the same framework that we applied to the description of testing moderator effects. In this section, we first review the steps for establishing mediation and then describe issues to consider in designing the study, analyzing the data, and interpreting the results. As was the case with moderation, we also note issues about which there appears to be confusion in counseling psychology research, on the basis of a review of studies published in *JCP* in 2001 that reported mediational analyses. Of the 54 articles that appeared in 2001, 10 (19%) contained a test of mediation or indirect effects.¹³ As before, the mediational analyses described in the identified studies are discussed on a general level so that particular studies and authors are not singled out. In the final section, we provide a step-by-step example to guide the reader through performing mediation analyses using multiple regression.

Guide to Testing Mediation Effects in Multiple Regression

According to MacKinnon, Lockwood, Hoffman, West, and Sheets (2002), the most common method for testing mediation in psychological research was developed by Kenny and his colleagues (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny, Kashy, & Bolger, 1998). According to this method, there are four steps (performed with three regression equations) in establishing that a variable (e.g., social support) mediates the relation between a predictor variable (e.g., counseling condition) and an outcome variable (e.g., well-being; see Figure 3A and Figure 3B). The first step is to show that there is a significant relation between the predictor and the outcome (see Path c in Figure 3A). The second step is to show that the predictor is related to the mediator (see Path a in Figure 3B). The third step is to show that the mediator (e.g., social support) is related to the outcome variable (e.g.,

¹² Aiken and West (1991, pp. 14–22) described the procedures for testing simple slopes when both the predictor and moderator are continuous variables.

¹³ The terms *mediated effects* and *indirect effects* are typically used interchangeably. According to MacKinnon et al. (2002), *mediation* is the more common term in psychology, whereas *indirect effect* comes from the sociological literature.

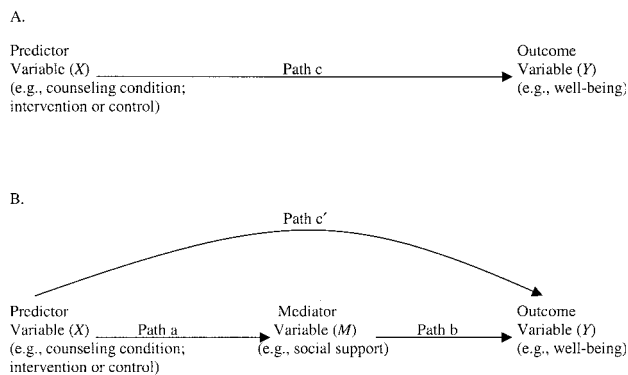


Figure 3. Diagram of paths in mediation models.

well-being). This is Path b in Figure 3B, and it is estimated controlling for the effects of the predictor on the outcome. The final step is to show that the strength of the relation between the predictor and the outcome is significantly reduced when the mediator is added to the model (compare Path c in Figure 3A with Path c' in Figure 3B). If social support is a complete mediator, the relation between counseling condition and well-being will not differ from zero after social support is included in the model. If social support is a partial mediator, which is more likely, the relation between counseling condition and well-being will be significantly smaller when social support is included but will still be greater than zero.

Designing a Study to Test Mediator Effects

In this section, we discuss four issues to consider in designing studies to test mediation: (a) the relation between the predictor and outcome variable, (b) choosing mediator variables, (c) establishing causation, and (d) factors that affect the power of the test of mediation.

Predictor–Outcome Relation

As mentioned, according to the model popularized by Kenny and colleagues (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny et al., 1998), the first step in the process of testing mediation is to establish that there is a significant relation between the predictor and outcome variable. That is, before one looks for variables that mediate an effect, there should be an effect to mediate. Therefore, in designing a mediational study, one generally should begin with predictor and outcome variables that are known to be significantly associated on the basis of prior research. As mentioned previously, the main purpose of mediational analyses is to examine why an association between a predictor and outcome exists.

There are, however, situations in which a researcher might want to look for evidence of mediation in the absence of a relation between a predictor and an outcome. In fact, Kenny et al. (1998) stated that this first step is not required (although a significant predictor–outcome relationship is implied if the predictor is related to the mediator and the mediator is related to the outcome). One example is a situation in which a treatment does not appear to be effective (i.e., no effect of predictor on outcome) because there

are multiple mediators producing inconsistent effects (Collins, Graham, & Flaherty, 1998; MacKinnon, 2000; MacKinnon, Krull, & Lockwood, 2000). For example, suppose an evaluation of a rape prevention program for men showed no differences between an intervention and a control group on an outcome measure of attitudes toward women. It may be that the intervention made men more empathic toward women, which was associated with positive changes in attitudes toward women. However, the intervention might also have made men more defensive, which might be associated with negative changes in attitudes toward women. The effects of these two mediators could cancel each other out, producing a nonsignificant intervention effect. In this case, it would be useful to perform mediational analyses in the absence of a predictor–outcome relation to identify these inconsistent mediators. Of course, to do such analyses these mediators would need to have been assessed, which often is not the case (MacKinnon, 1994). MacKinnon et al. (2001) provided an empirical example of an intervention with both positive and negative mediators.¹⁴

In a recent article, Shrout and Bolger (2002) recommended that inclusion of the first step in the Kenny model be based on whether the predictor is temporally distal or proximal to the outcome. Specifically, they recommended skipping the first step of the Kenny model in cases in which the predictor is distal to the outcome (such as in a long-term longitudinal study), because such studies often will lack power to detect the direct predictor–outcome relation. However, when the predictor is proximal to the outcome, or when theory suggests that the predictor–outcome relation is at least medium in size, they recommended retaining the first step in the Kenny model.

Choosing Mediator Variables

On a conceptual level, the proposed relations between the predictor and the mediator should be grounded in theory and clearly articulated. In other words, the rationale for the hypothesis that the predictor is related to or causes the mediator should have a clear theoretical rationale (see Holmbeck, 1997, for examples in which this rationale is lacking). Furthermore, given that the mediational model essentially is one in which the predictor causes the mediator, which in turn causes the outcome, the mediator ideally should be something that can be changed (MacKinnon et al., 2000).

Once potential mediators have been identified on theoretical grounds, there are practical issues to consider with regard to choosing specific mediators to test. In particular, the relations among the mediator, predictor, and outcome can affect the power of tests of mediation. For example, the power associated with the tests of the relations between the mediator and outcome (Path b in Figure 3B) and between the predictor and the outcome controlling for the mediator (Path c' in Figure 3B) decreases as the relation

¹⁴ A situation involving mediation but no significant predictor–outcome relation does not necessarily have to involve multiple mediators. This situation occurs more generally when the c path is opposite in sign to the ab path (Kenny et al., 1998). In this case, the mediator is a suppressor variable (see Cohen et al., 2003; MacKinnon et al., 2000; and Shrout & Bolger, 2002, for further discussion of suppression effects). Suppression occurs when the relation between a predictor and outcome becomes larger when the suppressor variable is included in the equation (as opposed to becoming smaller when a significant mediator is included in the equation).

between the predictor variable and the mediator increases (Kenny et al., 1998). That is, when more variance in the mediator is explained by the predictor, there is less variance in the mediator to contribute to the prediction of the outcome. Thus, as the predictor–mediator (Path a in Figure 3B) relation increases, a larger sample is needed to have the same amount of power to test the effects of Path b (mediator–outcome) and Path c' (predictor–outcome controlling for mediator), as would be the case if the relation between the predictor and mediator was smaller. Kenny et al. provided a formula to determine the “effective sample size” given the correlation between the predictor and the mediator: $N(1 - r_{xm}^2)$, where N is the sample size and r_{xm} is the correlation between the predictor and the mediator. For example, if your sample size is 900, and the predictor–mediator correlation is .30, the effective sample size is 819. However, if the predictor–mediator correlation is .70, the effective sample size is only 459. In other words, because of the high correlation between the predictor and mediator, power reduces to what it would be if your sample were 459 rather than 900 (thus, the sample size is effectively 459 rather than 900). Hoyle and Kenny (1999) presented the results of a simulation study demonstrating the effect of the size of the relation between the predictor and the mediator on the power of tests of mediation.

Another factor to consider in choosing mediators (from among theoretically viable candidates) is the size of the relation between the mediator and outcome (Path b in Figure 3B) relative to the size of the relation between the predictor and mediator (Path a in Figure 3B). According to Kenny et al. (1998), the relation between the mediator and outcome (Path b) and between the predictor and the mediator (Path a) should be comparable in size. However, Hoyle and Kenny (1999) noted that the power of tests of mediation is greatest when the relation between the mediator and the outcome (Path b) exceeds the relation between the predictor and the mediator (Path a). Thus, in choosing mediators, it is important to choose variables that are likely to have similar relations with the predictor and outcome variable (Path a = Path b), or somewhat stronger relations with the outcome than with the predictor (Path b > Path a), to maximize the power of the mediational test. These points were rarely, if ever, addressed in the *JCP* studies we reviewed, although they sometimes were an issue (e.g., correlations between predictors and mediators greater than .6 or stronger relations between predictors and mediators than between mediators and outcomes).

Once theoretically based mediators have been identified that satisfy the criteria just described with regard to their relations with the predictor and outcome, another factor to consider is the reliability of the measure of the mediator. Specifically, with lower reliability, the effect of the mediator on the outcome variable (Path b in Figure 3B) is underestimated, and the effect of the predictor variable on the outcome variable (Path c' in Figure 3B) is overestimated (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny et al., 1998). Thus, statistical analyses, such as multiple regression, that ignore measurement error underestimate mediation effects. Hoyle and Robinson (2003) have provided a formula for estimating the effects of unreliability on tests of mediation and recommend using a measure with a reliability of at least .90. They also describe four ways of modeling measurement error in SEM if such a highly reliable measure is not available and argue that the best approach is to use multiple measures and multiple measurement strategies. Most of the *JCP* mediation studies we reviewed used

multiple regression or used SEM programs to conduct a path analysis with single indicator variables (versus a model with latent variables). As a result, they did not take advantage of one of the primary reasons to use SEM (i.e., to model measurement error). This is important, because some mediators had either low (e.g., less than .70) or unreported reliability. None of the studies used multiple measurement strategies.

Establishing Causation

The process of mediation implies a causal chain; thus, definitions of mediation are almost always phrased in causal terms (see, e.g., Baron & Kenny, 1986; Hoyle & Smith, 1994; James & Brett, 1984; Judd & Kenny, 1981; Kenny et al., 1998; Kraemer et al., 2001). For example, Hoyle and Smith (1994) described a mediational hypothesis as “a secondary one that follows demonstration of an effect (assumed to be causal)” (p. 437; i.e., the predictor–outcome relation is causal). The mediator also is assumed to be caused by the predictor variable and to cause the outcome variable (Kenny et al., 1998). Consequently, the criteria for establishing causation need to be considered in study design.

The three primary criteria for establishing that one variable causes another are that (a) there is an association between the two variables (association), (b) the association is not spurious (isolation), and (c) the cause precedes the effect in time (direction; Hoyle & Smith, 1994; Menard, 1991). Satisfaction of these criteria can be seen as falling along a continuum, with one end of the continuum defined by nonexperimental correlational studies that merely establish an association between two variables and the other defined by experiments with random assignment to conditions. According to Wegener and Fabrigar (2000), even using a nonexperimental design, one can move farther along the continuum by controlling for the effects of other variables (isolation) or by collecting longitudinal data (direction; see Hoyle & Smith, 1994; Menard, 1991). Hoyle and Robinson (2003) have argued that the best approach is the “replicative strategy” in which all measures are administered at more than one point in time, which would require at least three assessments to assess mediation in a longitudinal study (Collins et al., 1998; Farrell, 1994). This is preferred over the sequential strategy in which the predictor, mediator, and outcome are measured at different points in time, because this design does not permit inferences of directionality.

With regard to the mediation studies published in *JCP* in 2001, all were nonexperimental, and little attention was paid to design features that would strengthen claims regarding causation. For example, only one study was longitudinal (but used a sequential rather than a replicative strategy), and only one controlled for a third variable that might affect the relation between the predictor and the outcome. Nonetheless, most researchers used causal language in describing their hypotheses and results.

Power

In the section on moderation, we noted that tests of interactions often have low power. The same is true of tests of mediation. We previously reviewed factors that can decrease the power of tests of mediation (e.g., high correlation between the mediator and the predictor). MacKinnon et al. (2002) recently performed a simulation study in which they compared the power of different methods

of testing mediation (see also Shrout & Bolger, 2002). The “causal steps” method described by Kenny (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny et al., 1998) was found to have adequate power only when sample sizes were large (greater than 500) or when the mediated effects were large. For example, the power to detect a medium effect with a sample size of 100 was only .28. The step requiring a significant effect of the predictor on the outcome (which we previously referred to as Step 1) led to the most Type II errors (i.e., lower power). Readers are encouraged to consult the MacKinnon et al. article for more information on alternative mediation tests. Hoyle and Kenny (1999) also performed a simulation study in which they examined the effects of several factors (e.g., reliability of the mediator) on the power of tests of mediation. Only samples of 200 had sufficient power (greater than .80). In designing studies, researchers need to estimate sample sizes a priori, using sources such as these, to ensure that the study has sufficient power. In the *JCP* studies reviewed, power was rarely mentioned, and most sample sizes were less than 200.

Analyzing the Data

Mediational analyses can be performed with either multiple regression or SEM. The logic of the analyses is the same in both cases. In general, SEM is considered the preferred method (Baron & Kenny, 1986; Hoyle & Smith, 1994; Judd & Kenny, 1981; Kenny et al., 1998). Some of the advantages of SEM are that it can control for measurement error, provides information on the degree of fit of the entire model, and is much more flexible than regression. For example, you can include multiple predictor variables, multiple outcome variables, and multiple mediators¹⁵ in the model as well as other potential causes of the mediator and outcome, including longitudinal data (Baron & Kenny, 1986; Hoyle & Smith, 1994; Judd & Kenny, 1981; MacKinnon, 2000; Quintana & Maxwell, 1999; Wegener & Fabrigar, 2000). However, in research areas in which it may be difficult to recruit a sufficiently large sample to perform SEM analyses (e.g., at least 200; see Quintana & Maxwell, 1999), it may be necessary to use multiple regression (Holmbeck, 1997). Furthermore, according to MacKinnon (2000), regression is the most common method for testing mediation (see Hoyle & Kenny, 1999, for a simulation study comparing regression with SEM for testing mediation). Therefore, we first describe methods for testing mediation using regression and then describe methods using SEM.

As mentioned, the method outlined by Kenny (e.g., Baron & Kenny, 1986; Kenny et al., 1998) is the most commonly used approach in the psychological literature. Using multiple regression, this approach involves testing three equations. First, the outcome variable is regressed on the predictor to establish that there is an effect to mediate (see Path c in Figure 3A). Second, the mediator is regressed on the predictor variable to establish Path a (see Figure 3B) in the mediational chain. In the third equation, the outcome variable is regressed on both the predictor and the mediator. This provides a test of whether the mediator is related to the outcome (Path b) as well as an estimate of the relation between the predictor and the outcome controlling for the mediator (Path c'). If the relation between the predictor and the outcome controlling for the mediator is zero, the data are consistent with a complete mediation model (i.e., the mediator completely accounts for the relation between the predictor and outcome). If the relation be-

tween the predictor and the outcome is significantly smaller when the mediator is in the equation (Path c') than when the mediator is not in the equation (Path c), but still greater than zero, the data suggest partial mediation. However, it is not enough to show that the relation between the predictor and outcome is smaller or no longer is significant when the mediator is added to the model. Rather, one of several methods for testing the significance of the mediated effect should be used (see MacKinnon et al., 2002, for a comparison of several different methods and Shrout & Bolger, 2002, for an alternative bootstrapping procedure). MacKinnon et al.'s review of published studies indicated that the majority did not test the significance of the mediating variable effect. This also was true of the mediation studies published in *JCP* in 2001.

The method described by Kenny et al. (1998) to test the significance of the mediated effect is as follows: Because the difference between the total effect of the predictor on the outcome (Path c in Figure 3A) and the direct effect of the predictor on the outcome (Path c' in Figure 3B) is equal to the product of the paths from the predictor to the mediator (Path a) and from the mediator to the outcome (Path b), the significance of the difference between Paths c and c' can be assessed by testing the significance of the products of Paths a and b. Specifically, the product of Paths a and b is divided by a standard error term. The mediated effect divided by its standard error yields a z score of the mediated effect. If the z score is greater than 1.96, the effect is significant at the .05 level. The error term used by Kenny and colleagues (Baron & Kenny, 1986; Kenny et al., 1998) is the square root of $b^2sa^2 + a^2sb^2 + sa^2sb^2$, where a and b are unstandardized regression coefficients and sa and sb are their standard errors. Note that this differs from Sobel's (1982) test, which is the most commonly used standard error. Sobel's test does not include the last term (sa^2sb^2), which typically is small. These two methods performed very similarly in MacKinnon et al.'s (2002) simulation study.

Although we are focusing here on the use of multiple regression, as mentioned, there are various ways to test mediational models in SEM. Holmbeck (1997) described a strategy for SEM that is virtually identical to that used with regression (i.e., testing the fit of the predictor–outcome model and the fit of the predictor–mediator–outcome model, as well as the predictor–mediator and mediator–outcome paths). These analyses provide tests of Steps 1 through 3 outlined previously. To test the significance of the mediated effect, the fit of the predictor–mediator–outcome model is compared with and without the direct path from the predictor and outcome constrained to zero. A mediational model is supported if the model with the direct path between the predictor and outcome does not provide a better fit to the data (i.e., the direct path between the predictor and outcome is not significant). Hoyle and Smith (1994) described a somewhat simpler approach in which the predictor–outcome path is compared in models with and without the mediator. As in regression, if the predictor–outcome

¹⁵ Procedures for assessing multiple mediators in regression have been described by MacKinnon (2000). Cohen et al. (2003, pp. 460–467) also described methods for calculating total, direct, indirect, and spurious effects for multiple variables using multiple regression. In addition, several authors have provided detailed accounts of testing multiple mediator models in SEM (e.g., Brown, 1997; MacKinnon, 2000; MacKinnon et al., 2001).

path is zero with the mediator in the model, there is evidence of complete mediation. The significance of the mediated effect also can be obtained in a single model by multiplying the coefficients for Paths a (predictor to mediator) and b (mediator to outcome). Tests of the significance of indirect effects are available in most SEM programs (see Brown, 1997, for a description of testing mediation models in LISREL). Few of the *JCP* studies reviewed that assessed mediated or indirect effects used the Kenny framework either in regression or SEM. Some compared models with and without direct paths from the predictor to the outcome, and some only reported the significance of the indirect effects.

Interpreting the Results

If a researcher has found support for all of the conditions for mediation mentioned earlier, what conclusions are appropriate? Next, we briefly describe three factors to consider when interpreting the results of mediation analyses.

Alternative Equivalent Models

One issue that must be acknowledged when interpreting the results of mediational analyses is that, even if the four conditions mentioned earlier are met, there are likely to be other models that are consistent with the data that also are correct (Kenny et al., 1998; Quintana & Maxwell, 1999). MacCallum, Wegener, Uchino, and Fabrigar (1993) provided a complete discussion of this issue, and we encourage readers to refer to their analysis. Briefly, they showed that for any given model there generally are alternative models with different patterns of relations among variables that fit the data as well as the original model. According to their review of 53 published studies that used SEM, approximately 87% of the models had alternative equivalent models, with the average being 12. MacCallum et al. provided rules for calculating the number of alternative equivalent models, although some SEM programs (e.g., AMOS) have options that will generate all of the possible models from the observed variables and calculate the percentage of the possible models whose fit is better than, worse than, or comparable to the original model. Of course, some of these models can be rejected on the basis of the meaningfulness of the model, and MacCallum et al. reviewed design factors that affect the meaningfulness of alternative models. For example, paths to experimentally manipulated variables would not be meaningful (e.g., that counseling outcomes cause assignment to treatment condition), nor would paths that move backward in time in longitudinal studies. Thus, the number of alternative models is greater when the data are cross-sectional and correlational.

Even with experimental or longitudinal data, there are likely to be alternative models that fit the data, and researchers are encouraged to identify and test such models. However, most researchers in the studies reviewed by MacCallum et al. (1993) and in the studies that we reviewed in *JCP* asserted the validity of their model on the basis of goodness-of-fit indexes without acknowledging or testing any alternatives. The existence of equivalent models “presents a serious challenge to the inferences typically made by researchers” (MacCallum et al., 1993, p. 196).

Omitted Variables

Another problem that may affect the interpretation of mediational analyses is omitted variables. Specifically, mediational anal-

yses may yield biased estimates if variables that cause both the mediator and the outcome are not included in the model, because the association between the mediator and the outcome may be due to third variables that cause both (James & Brett, 1984; Kenny et al., 1998). Judd and Kenny (1981) provided a detailed example in which adding a variable that is related to both the mediator and the outcome substantially changes the results of the mediational analysis. Although this is a difficult problem to solve, there are ways to address it. For example, common causes of the mediator and the outcome, such as social desirability, can be included directly in the model. In addition, using different methods (e.g., self-reports and peer ratings) to measure the mediator and outcome will reduce the extent to which they are correlated because of common method variance.

Causation

We discussed causation under the section on study design, but this topic also is relevant to interpreting the results of mediation analyses. All of the studies testing mediation that we reviewed in *JCP* were nonexperimental, as was true of most of the studies using SEM reviewed by MacKinnon et al. (2002). In the *JCP* studies, causal language often was used even though causal inferences generally cannot be made on the basis of nonexperimental data (Cohen et al., 2003; Hoyle & Smith, 1994; Kraemer et al., 2001). James and Brett (1984) recommended that researchers attend to all conditions necessary for establishing causation before conducting mediational tests and using these tests to support causal inferences. If one or more sources of specification error is viable (e.g., misspecification of causal direction or an unmeasured variable problem), exploratory procedures should be used and interpreted only in correlational terms (e.g., the correlation between the predictor and outcome is diminished if the mediator is controlled).¹⁶ However, in this case the mediator cannot be said to explain how the predictor and outcome are related, which essentially defeats the purpose of testing a mediational model. Others take a somewhat more liberal stance, arguing that, with correlational data, all that can be said is that the causal model is consistent with the data (Kraemer et al., 2001). In this case, it must be acknowledged that other models also are consistent with the data, as discussed previously.

Conclusions

There are many issues to consider when designing, conducting, and interpreting mediational analyses. We acknowledge that it is not possible for every consideration mentioned to be addressed in every study. However, according to our review of mediational research published in *JCP*, there definitely is room for improvement. For example, virtually all of the mediational analyses in the studies we reviewed were performed with cross-sectional correlational data. Very few attempts were made to control for common causes of the mediator and the outcome. The direction of the

¹⁶ In this example, a mediator is very similar to a confounding variable, which is a variable that distorts the relation between two other variables (MacKinnon et al., 2000). For example, once the effect of age is controlled, income is no longer related to cancer prevalence. Unlike mediation, confounding does not imply causality.

relations among variables often was unclear. Nonetheless, authors typically discussed results using causal language. Authors sometimes acknowledged that no causal conclusions could be drawn, even though they used causal language. In most of the studies we reviewed, the authors concluded that their model fit the data without acknowledging or testing alternative models. However, the best evidence for mediation requires showing not only that the data are consistent with the proposed mediation model but also that other models are either theoretically implausible or inconsistent with the data (Smith, 2000). If a compelling argument cannot be made for the superiority of one model over another, additional research needs to be conducted to distinguish among the alternative models (MacCallum et al., 1993).

Example: Testing Mediation Using Multiple Regression

To illustrate how mediator effects may be investigated with multiple regression, we again provide a step-by-step example. As mentioned, in this case we use actual data to illustrate issues that arise when using real, versus simulated, data.

Designing the Study

The data we are using to illustrate the process of conducting mediational analyses with multiple regression were collected by Patricia A. Frazier from 894 women who responded to a random-digit dialing telephone survey regarding traumatic experiences and posttraumatic stress disorder (PTSD). Participants were asked whether they had experienced several traumatic events and, if so, to indicate which was their worst lifetime trauma. These events had occurred an average of 10 years previously. One finding from this study was that individuals whose self-nominated worst event happened directly to them (e.g., sexual assault) reported more current symptoms of PTSD than those whose worst events did not happen directly to them (e.g., life-threatening illness of a close friend or family member). Although this may seem obvious, the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) does not distinguish between directly and indirectly experienced events in the stressor criterion for PTSD. Because event type (directly vs. indirectly experienced) was associated with PTSD symptoms (i.e., signifi-

cant predictor–outcome relationship), examining mediators of this relationship can help us to understand why directly experienced events are more likely to lead to PTSD than are indirectly experienced events.

The mediator we chose to examine was self-blame. We chose self-blame as a potential mediator because previous theory and research suggest that it is one of the strongest correlates of post-traumatic distress (e.g., Weaver & Clum, 1995). It also seemed that individuals would be more likely to blame themselves for events that happened directly to them (e.g., an accident or a sexual assault) than for events that happened to others (e.g., life-threatening illness of a close friend). Other possible mediators were rejected because, although they might be associated with higher levels of PTSD symptoms, they were unlikely to have been caused by the experience of a “direct” trauma. For example, individuals who have experienced more lifetime traumas report more symptoms of PTSD, but it seemed unlikely that experiencing a direct trauma would cause one to experience more lifetime traumas. In addition, unlike past traumas, self-blame is a factor that can be changed. Thus, the mediational hypothesis we tested was that individuals will blame themselves more for directly experienced events (Path a), and individuals who engage in more self-blame will report more PTSD symptoms (Path b). Finally, we hypothesized that once the relation between self-blame and PTSD symptoms was accounted for, there would be a weaker relation between event type (directly vs. indirectly experienced events) and PTSD (i.e., Path c' will be smaller than Path c). Thus, self-blame was hypothesized to be a partial (vs. complete) mediator. Given our sample size ($N = 894$), we had sufficient power to detect medium to large mediated effects (MacKinnon et al., 2002).

Analyzing the Data

Table 2 contains the analyses necessary to examine this mediational hypothesis. Following the steps outlined earlier for testing mediation, we first established that event type (the predictor) was related to PTSD symptoms (the outcome) by regressing PTSD symptoms on the event-type variable (Step 1). The unstandardized regression coefficient ($B = 1.32$) associated with the effect of event type on number of PTSD symptoms was significant ($p <$

Table 2
Testing Mediator Effects Using Multiple Regression

Testing steps in mediation model	<i>B</i>	<i>SE B</i>	95% CI	β
Testing Step 1 (Path c) Outcome: current PTSD symptoms Predictor: event type (direct vs. indirect) ^a	1.32	0.22	0.90, 1.75	.21**
Testing Step 2 (Path a) Outcome: self-blame Predictor: event type	0.50	0.04	0.41, 0.58	.38**
Testing Step 3 (Paths b and c') Outcome: current PTSD symptoms Mediator: self-blame (Path b) Predictor: event type	0.95 0.86	0.18 0.23	0.60, 1.29 0.41, 1.31	.19** .13**

Note. CI = confidence interval; PTSD = posttraumatic stress disorder.

^a 0 = indirectly experienced trauma, 1 = directly experienced trauma.

** $p < .001$.

.0001). Thus, Path *c* was significant, and the requirement for mediation in Step 1 was met. To establish that event type was related to self-blame (the hypothesized mediator), we regressed self-blame on the event type variable (Step 2). The unstandardized regression coefficient ($B = 0.50$) associated with this relation also was significant at the $p < .0001$ level, and thus the condition for Step 2 was met (Path *a* was significant). To test whether self-blame was related to PTSD symptoms, we regressed PTSD symptoms simultaneously on both self-blame and the event type variable (Step 3). The coefficient associated with the relation between self-blame and PTSD (controlling for event type) also was significant ($B = 0.95$, $p < .0001$). Thus, the condition for Step 3 was met (Path *b* was significant). This third regression equation also provided an estimate of Path *c'*, the relation between event type and PTSD, controlling for self-blame. When that path is zero, there is complete mediation. However, Path *c'* was 0.86 and still significant ($p < .001$), although it was smaller than Path *c* (which was 1.32).

There are several ways to assess whether this drop from 1.32 to 0.86 (i.e., from *c* to *c'*) is significant. Because $c - c'$ is equal to the product of Paths *a* and *b*, the significance of the difference between *c* and *c'* can be estimated by testing the significance of the products of Paths *a* and *b*. Specifically, you divide the product of Paths *a* and *b* by a standard error term. Although there are several different ways to calculate this standard error term, we used the error term used by Kenny and colleagues (Baron & Kenny, 1986; Kenny et al., 1998) described earlier: the square root of $b^2sa^2 + a^2sb^2 + sa^2sb^2$, where *a* and *b* are unstandardized regression coefficients and *sa* and *sb* are their standard errors. We used this term because it is likely to be more familiar to readers through Kenny's writings. To reiterate, the mediated effect divided by its standard error yields a *z* score of the mediated effect. If the *z* score is greater than 1.96, the effect is significant at the .05 level. In our case, we multiplied the unstandardized regression coefficient weights for Path *a* (0.50) and Path *b* (0.95) and divided by the square root of $(0.90)(0.002) + (0.25)(0.03) + (0.002)(0.03)$, which yielded $0.475/0.097 = 4.90$. Thus, self-blame was a significant mediator even though the *c'* path was significant.¹⁷ Shrout and Bolger (2002) also recommended calculating the confidence interval around the estimate of the indirect effect. The formula for calculating a 95% confidence interval is the product of Paths *a* and *b* $\pm s_{ab} z_{.975}$, where $z_{.975}$ is equal to the constant 1.96 and s_{ab} is the standard error term calculated earlier. For our example, the 95% confidence interval would be $0.475 \pm 0.097 (1.96) = 0.29$ to 0.67. This confidence interval does not include zero, which is consistent with the conclusion that there is mediation (i.e., the indirect effect is not zero).

Another way to describe the amount of mediation is in terms of the proportion of the total effect that is mediated, which is defined by ab/c (Shrout & Bolger, 2002). Using the unstandardized regression coefficients from our example, we get $0.475/1.32 = .36$. Thus, about 36% of the total effect of event type on PTSD symptoms is mediated by self-blame. However, a sample size of at least 500 is needed for accurate point and variance estimates of the proportion of total effect mediated (MacKinnon, Warsi, & Dwyer, 1995). Also, it is important to note that this is just a way of describing the amount of mediation rather than a test of the significance of the mediated effect.

Interpreting the Results

What can we conclude from this test of our hypothesis that self-blame partially mediates the relation between type of trauma experienced and PTSD symptoms (i.e., that directly experienced events result in more PTSD because they lead to more self-blame, which in turn leads to more PTSD symptoms)? In terms of causation, a strong argument can be made that the traumatic event (the predictor) preceded both self-blame (the mediator) and PTSD (the outcome). However, it could be the case that individuals who are suffering from more PTSD symptoms are more likely to blame themselves (i.e., that the outcome causes the mediator). In fact, when we tested this alternative model, PTSD also was a significant mediator of the relation between event type and self-blame. Thus, there are alternative models that are consistent with the data. We also did not control for other factors that may be related to or cause both self-blame and PTSD, such as the personality trait neuroticism. Thus, all we can say at this point is that our data are consistent with models in which self-blame causes PTSD and PTSD causes self-blame. We also must acknowledge that the mediational relations we found might not have been evident if other variables that cause both self-blame and PTSD had been included in the model.

How does this example compare with the design considerations mentioned before? First, we began by establishing that there was a significant predictor–outcome relation. We next established a theoretical rationale for why the predictor variable would be related to the mediator self-blame and chose a mediator that potentially is alterable. Ideally, these decisions regarding potential mediator variables are made before data collection. The Path *a* relationship between the predictor and the mediator ($\beta = .38$) was not so high that multicollinearity would be a problem. However, the Path *a* relation between the predictor and mediator ($\beta = .38$) was larger than the Path *b* relation between the mediator and the outcome ($\beta = .19$); our power would have been greater if Path *b* were equal to or larger than Path *a*. In addition, our measure of self-blame was not without measurement error, which reduces the power of the test of the mediated effect. With regard to the fourth step, even though the relation between the predictor and outcome remained significant after the mediator had been controlled, a test of the mediated effect revealed that there was significant mediation. Nonetheless, given measurement error in the mediator and the fact that Path *a* was larger than Path *b*, both of which reduce the power of the test of mediation, we may have underestimated the extent to which self-blame mediated the relation between event type and PTSD.

After finding data consistent with a mediational model, what comes next? First, given that there are alternative models that may fit the data equally well, it is important to conduct additional studies that can help rule out these alternative models. Experimentally manipulating the predictor or the outcome would help to rule out alternative models in which the mediator causes the predictor or the outcome causes the mediator. However, experimental manipulation would not be ethical in cases like our example. As

¹⁷ An online calculator for this test is available that provides tests of mediation using three different error terms (Preacher & Leonardelli, 2003). In our example, the mediated effect is significant regardless of which error term is used.

described earlier, alternative models also can be tested in nonexperimental studies that incorporate design features that might rule out alternative models (e.g., by collecting longitudinal data, including other mediators, measuring common causes of the mediator and outcome, and using multiple measurement strategies). If plausible alternative models are rejected, a mediator might suggest areas of intervention (Baron & Kenny, 1986). For example, if we find stronger evidence that self-blame causes PTSD (rather than the other way around), we might want to develop an intervention to decrease self-blame and thereby reduce PTSD symptoms.

CONCLUSION

In summary, we have offered counseling researchers step-by-step guides to testing mediator and moderator effects that can be used both in planning their own research and in evaluating published research and have provided an example of each type of analysis. We hope that these guides will increase the extent to which counseling researchers move beyond the testing of direct effects and include mediators and moderators in their analyses and improve the quality of those tests when they are conducted. Toward these ends, the Appendix contains checklists to use in designing and evaluating tests of moderator and mediator effects.

ADDITIONAL RESOURCES

Although we have presented additional resources throughout the article, we want to highlight a few of these resources. As mentioned previously, it is important for researchers to consult these primary sources (and new sources as they emerge) to gain a better understanding of the underlying statistical and conceptual issues in testing moderation and mediation. Helpful sources that cover both mediation and moderation are the classic article by Baron and Kenny (1986) and the new regression textbook by Cohen et al. (2003). Particularly helpful sources for further information on moderation include Aiken and West (1991), Jaccard et al. (1990), and West et al. (1996). Useful sources for further information on mediation include Kenny et al. (1998), Shrout and Bolger (2002), and MacKinnon et al. (2002). Finally, MacKinnon (2003) and Kenny (2003) both have Web sites to which one can submit questions regarding mediation. A user-friendly guide to testing interactions using multiple regression also can be found on the World Wide Web (Preacher & Rucker, 2003).

References

- Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management Research*, *21*, 1141–1158.
- Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods*, *4*, 291–323.
- Aguinis, H., Bommer, W. H., & Pierce, C. A. (1996). Improving the estimation of moderating effects by using computer-administered questionnaires. *Educational and Psychological Measurement*, *56*, 1043–1047.
- Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods*, *2*, 315–339.
- Aguinis, H., & Pierce, C. A. (1998). Statistical power computations for detecting dichotomous moderator variables with moderated multiple regression. *Educational and Psychological Measurement*, *58*, 668–676.
- Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, *82*, 192–206.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*, 308–314.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.
- Bissonnette, V., Ickes, W., Bernstein, I., & Knowles, E. (1990). Personality moderating variables: A warning about statistical artifact and a comparison of analytic techniques. *Journal of Personality*, *58*, 567–587.
- Bollen, K. A., & Paxton, P. (1998). Interactions of latent variables in structural equation models. *Structural Equation Modeling*, *5*, 267–293.
- Brown, R. L. (1997). Assessing specific mediational effects in complex theoretical models. *Structural Equation Modeling*, *4*, 142–156.
- Busemeyer, J., & Jones, L. R. (1983). Analysis of multiplicative causal rules when the causal variables are measured with error. *Psychological Bulletin*, *93*, 549–562.
- Chaplin, W. F. (1991). The next generation in moderation research in personality psychology. *Journal of Personality*, *59*, 143–178.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Collins, L. M., Graham, J. W., & Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, *33*, 295–312.
- Coming, A. F. (2002). Self-esteem as a moderator between perceived discrimination and psychological distress among women. *Journal of Counseling Psychology*, *49*, 117–126.
- Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin*, *102*, 414–417.
- Cross, S. E., & Madson, L. (1997). Models of the self: Self-construals and gender. *Psychological Bulletin*, *122*, 5–37.
- DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods*, *1*, 261–277.
- Dunlap, W. P., & Kemery, E. R. (1987). Failure to detect moderating effects: Is multicollinearity the problem? *Psychological Bulletin*, *102*, 418–420.
- Farrell, A. D. (1994). Structural equation modeling with longitudinal data: Strategies for examining group differences and reciprocal relationships. *Journal of Consulting and Clinical Psychology*, *62*, 477–487.
- Frazier, P., Tix, A., & Barnett, C. L. (2003). The relational context of social support. *Personality and Social Psychology Bulletin*, *29*, 1113–1146.
- Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science*, *26*, 797–833.
- Grissom, R. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*, 155–165.
- Holmbeck, G. N. (1997). Toward terminological, conceptual, and statisti-

- cal clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literatures. *Journal of Consulting and Clinical Psychology*, 65, 599–610.
- Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks, CA: Sage.
- Hoyle, R. H., & Robinson, J. I. (2003). Mediated and moderated effects in social psychological research: Measurement, design, and analysis issues. In C. Sansone, C. Morf, & A. T. Panter (Eds.), *Handbook of methods in social psychology*. Thousand Oaks, CA: Sage.
- Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural models: A conceptual overview. *Journal of Consulting and Clinical Psychology*, 62, 429–440.
- Jaccard, J., Turrissi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, 117, 348–357.
- Jaccard, J., & Wan, C. K. (1996). LISREL approaches to interaction effects in multiple regression. Thousand Oaks, CA: Sage.
- James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69, 307–321.
- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5, 602–619.
- Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology*, 46, 433–465.
- Kenny, D. (2003). *Mediation*. Retrieved November 10, 2003, from <http://users.rcn.com/dakenny/mediate.htm>
- Kenny, D. A., & Judd, C. M. (1984). Estimating the linear and interactive effects of latent variables. *Psychological Bulletin*, 105, 361–373.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 233–265). New York: Oxford University Press.
- Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, 158, 848–856.
- Lakey, B., & Drew, J. B. (1997). A social-cognitive perspective on social support. In G. R. Pierce, B. Lakey, I. G. Sarason, & B. R. Sarason (Eds.), *Sourcebook of social support and personality* (pp. 107–140). New York: Plenum Press.
- Lee, R. M., Draper, M., & Lee, S. (2001). Social connectedness, dysfunctional interpersonal behaviors, and psychological distress: Testing a mediator model. *Journal of Counseling Psychology*, 48, 310–318.
- Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious “moderator effects”: Illustrated substantively with the hypothesized (“synergistic”) relation between spatial and mathematical ability. *Psychological Bulletin*, 107, 385–393.
- MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, 118, 405–421.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40.
- MacKinnon, D. P. (1994). Analysis of mediating variables in prevention and intervention research. In A. Cazares & L. A. Beatty (Eds.), *Scientific methods for prevention intervention research* (NIDA Research Monograph 139, DHHS Publication No. 94-3631, pp. 127–153). Washington, DC: U.S. Government Printing Office.
- MacKinnon, D. P. (2000). Contrasts in multiple mediator models. In J. S. Rose, L. Chassin, C. C. Presson, & S. J. Sherman (Eds.), *Multivariate applications in substance use research: New methods for new questions* (pp. 141–160). Mahwah, NJ: Erlbaum.
- MacKinnon, D. (2003). *Mediation*. Retrieved November 10, 2003, from <http://www.public.asu.edu/~davidpm/ripl/mediate.htm>
- MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, 17, 144–158.
- MacKinnon, D. P., Goldberg, L., Clarke, G. N., Elliot, D. L., Cheong, J., Lapin, A., et al. (2001). Mediating mechanisms in a program to reduce intentions to use anabolic steroids and improve exercise self-efficacy and dietary behavior. *Prevention Science*, 2, 15–27.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. (2000). Mediation, confounding, and suppression: Different names for the same effect. *Prevention Science*, 1, 173–181.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, 30, 41–62.
- Marsh, H. W. (2002, April). *Structural equation models of latent interactions: Evaluation of alternative strategies*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Mason, C. A., Tu, S., & Cauce, A. M. (1996). Assessing moderator variables: Two computer simulation studies. *Educational and Psychological Measurement*, 56, 45–62.
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113, 181–190.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376–390.
- Menard, S. (1991). *Longitudinal research: Quantitative applications in the social sciences*. Newbury Park, CA: Sage.
- Moulder, B. C., & Algina, J. (2002). Comparison of methods for estimating and testing latent variable interactions. *Structural Equation Modeling*, 9, 1–19.
- Norcross, J. (2001). Purposes, processes, and products of the task force on empirically supported therapy relationships. *Psychotherapy*, 38, 345–356.
- Overton, R. C. (2001). Moderated multiple regression for interactions involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological Methods*, 6, 218–233.
- Ping, R. A., Jr. (1996). Latent variable interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin*, 119, 166–175.
- Preacher, K., & Rucker, D. (2003). *A primer on interaction effects in multiple linear regression*. Retrieved November 10, 2003, from <http://www.unc.edu/~preacher/lcamlm/interactions.htm>
- Preacher, K., & Leonardelli, G. (2003). *Calculation for the Sobel test: An interactive calculation tool for mediation tests*. Retrieved November 10, 2003, from <http://www.unc.edu/~preacher/sobel/sobel.htm>
- Quintana, S. M., & Maxwell, S. E. (1999). Implications of recent developments in structural equation modeling for counseling psychology. *The Counseling Psychologist*, 27, 485–527.
- Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology*, 77, 336–342.
- Schumacker, R., & Marcoulides, G. (Eds.). (1998). *Interaction and non-linear effects in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and non-

- experimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Smith, E. R. (2000). Research design. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 17–39). New York: Cambridge University Press.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology 1982* (pp. 290–312). Washington, DC: American Sociological Association.
- Stone-Romero, E. F., Alliger, G., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management*, 20, 167–178.
- Stone-Romero, E. F., & Anderson, L. E. (1994). Techniques for detecting moderating effects: Relative statistical power of multiple regression and the comparison of subgroup-based correlation coefficients. *Journal of Applied Psychology*, 79, 354–359.
- Weaver, T., & Clum, G. (1995). Psychological distress associated with interpersonal violence: A meta-analysis. *Clinical Psychology Review*, 15, 115–140.
- Wegener, D., & Fabrigar, L. (2000). Analysis and design for nonexperimental data addressing causal and noncausal hypotheses. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 412–450). New York: Cambridge University Press.
- West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, 64, 1–49.

Appendix

Checklists for Evaluating Moderation and Mediation Analyses

Checklist for Evaluating Moderator Analyses Using Multiple Regression

- Was a strong theoretical rationale for the interaction provided? Was the specific form of the interaction specified?
- Was power calculated a priori? If power was low, was this mentioned as a limitation?
- Was the size of the interaction considered a priori to determine needed sample size?
- Was the overall effect size considered a priori to determine needed sample size?
- If there was a categorical predictor or moderator, were the sample sizes for each group relatively equal?
- Was the assumption of homogeneous error variance checked if categorical variables were used? If the assumption was violated, were alternative tests used?
- Were continuous variables sufficiently reliable (e.g., above .80)?
- Were continuous variables normally distributed (i.e., no range restriction)?
- Was the outcome variable sensitive enough to capture the interaction?
- Were regression procedures used whether variables were categorical or continuous? Were continuous variables kept continuous?
- If there were categorical variables, was the coding scheme used appropriate for the research questions?
- Were continuous variables centered or standardized?
- Was the interaction term created correctly (i.e., by multiplying the predictor and moderator variables)?
- Was the equation structured correctly (i.e., predictor and moderator entered before the interaction term)?
- Were first-order effects interpreted correctly given the coding system used?
- Were unstandardized (rather than standardized) coefficients interpreted?
- Was the significance of the interaction term assessed appropriately (i.e., by examining the change in R^2 associated with the interaction term)?
- Was the interaction plotted if significant?
- Were the simple slopes compared?
- If covariates were used, were interactions with other terms in the model assessed?

- If multiple moderators were tested, was Type I error addressed?
- Did the interpretation of the results reflect the actual effect size of the interaction?

Checklist for Evaluating Mediation Analyses Using Multiple Regression

- Was the predictor significantly related to the outcome? If not, was there a convincing rationale for examining mediation?
- Was there a theoretical rationale for the hypothesis that the predictor causes the mediator? Was the mediator something that can be changed?
- What is the “effective sample size” given the correlation between the predictor and mediator? That is, was the relation between the predictor and mediator so high as to compromise power?
- Was the relation between the predictor and the outcome (Path b) greater than or equal to the relation between the predictor and the mediator (Path a)?
- Were the mediators adequately reliable (e.g., $\alpha > .90$)?
- Was unreliability in the mediators (e.g., $\alpha < .70$) addressed through tests that estimate the effects of unreliability or the use of SEM?
- To what extent did the design of the study enable causal inferences?
- Was power mentioned either as an a priori consideration or as a limitation?
- Were all four steps in establishing mediation addressed in the statistical analyses?
- Was the significance of the mediation effect formally tested?
- Were alternative equivalent models acknowledged or tested?
- Were variables that seem likely to cause both the mediator and the outcome included in analyses, or were multiple measurement methods used?
- Did the study design allow for the type of causal language used in the interpretation of results?

Received March 18, 2003
 Revision received July 1, 2003
 Accepted July 3, 2003 ■