Rehabilitation Psychology
2008, Vol. 53, No. 3, 321–339

0090-5550/08/$12.00 DOI: 10.1037/a0013021

Multiple Regression and Correlation Techniques: Recent Controversies and Best Practices

William T. Hoyt, Zac E. Imel, and Fong Chan
University of Wisconsin—Madison

Objective: This article presumes familiarity with the basics of multiple regression and correlation (MRC) methods and addresses recent controversies and emerging innovations. Areas of emphasis include linking analyses to theory-driven hypotheses, treatment of covariates in hierarchical regression models, recent debates about the testing of mediator and moderator hypotheses, and incorporating confidence intervals into reports of findings using MRC. ***Conclusions:*** Two important conceptual innovations (linking analyses closely to theory-derived hypotheses; focusing interpretations on effect sizes and confidence intervals rather than *p* values) can increase the scientific yield for researchers making use of MRC methods in rehabilitation psychology.

Keywords: data analysis, multiple regression, research methods, confidence intervals, mediation

Multiple regression and correlation (MRC) analyses provide a flexible data-analytic framework for addressing a wide variety of questions of interest to rehabilitation psychologists. Regression models can accommodate multiple correlated predictor variables, including nominal (categorical) variables, and can be used to test sophisticated models involving mediation or moderation (statistical interactions). They can be used to statistically control for confounding variables and to examine the predictive power of sets of predictor variables as well as the unique association of a single predictor with the dependent variable (DV).

Regression methods have been popular with rehabilitation researchers. Examination of approximately 200 articles published in *Rehabilitation Psychology* between February 2004 and February 2008 revealed that more than one third of these studies used some form of regression analysis. An additional reason why familiarity with regression methods is valuable is that these techniques form the foundation for multivariate methods such as factor analysis, structural equation modeling, and multilevel modeling. Familiarity with analysis and interpretation issues in MRC is therefore important for consumers and users of these more sophisticated methods.

One of the authors of the present article recently collaborated on an introduction to MRC directed at researchers in rehabilitation counseling and rehabilitation psychology (Hoyt, Leierer, & Millington, 2006). That article defined basic terms and notational conventions and offered guidance about fundamental issues such as the choice between standardized and unstandardized regression coefficients, interpretation of partial regression coefficients (i.e., regression coefficients for a given predictor variable when other predictors are also in the regression equation), power analysis, and factors affecting the magnitude of correlation and regression coefficients. The goal of the present article is to build on this foundation, reviewing guidelines for addressing more sophisticated research hypotheses (such as those involving mediation or moderation) and providing illustrative examples for reporting and interpreting findings. We review recent controversies regarding definitions and analytical methods and provide recommendations to assist both authors and readers with design, analysis, and interpretation using MRC.

General Framework for MRC

Multiple regression analyses simultaneously examine the association between multiple predictor variables (X_1, X_2, X_3, etc.) and a single criterion variable (Y). Relationships among the variables are summarized in a regression equation. For the two-predictor case, this equation takes the form

$$\hat{Y}_i = B_1X_{1i} + B_2X_{2i} + B_0, \tag{1}$$

where $\hat{Y}_i$ represents the predicted score on the criterion variable for person *i,* to be computed from that person's known scores on X_1 and X_2. The regression coefficients B_1 and B_2 are the multipliers for X_{1i} and X_{2i}, respectively. (For simplicity, we generally omit the *i* subscripts in the text.) The regression coefficients are chosen so as to maximize the proportion of *Y* variance explained by the linear composite on the right side of Equation 1 or, identically, to minimize the errors of prediction, notated $Y_i - \hat{Y}_i$. The third regression coefficient (B_0) is called the *constant* or the *intercept* and denotes the predicted value of *Y* for a person with scores $X_1 = X_2 = 0$.

When this equation is estimated from sample data, the values of the regression coefficients (and their statistical significance) are

Editor's Note. Kathleen Chwalisz served as the action editor for this article.—TRE

William T. Hoyt and Zac E. Imel, Department of Counseling Psychology, University of Wisconsin—Madison; Fong Chan, Department of Rehabilitation Psychology and Special Education, University of Wisconsin—Madison.

Correspondence concerning this article should be addressed to William T. Hoyt, Department of Counseling Psychology, University of Wisconsin—Madison, 321 Education Building, 1000 Bascom Mall, Madison, WI 53706-1398. E-mail: wthoyt@wisc.edu

informative about the strength of association between each independent variable (IV) and the DV. B_1 is the unstandardized regression coefficient because it carries the original units of X_1. Because both X_1 and X_2 are predictors in this equation, B_1 is an estimate of the predicted change in Y for a one-unit change in X_1 when X_2 is held constant (i.e., statistically controlled). In causal terms, B_1 may be interpreted as the unique effect of X_1 on Y, controlling for X_2.

Because X_1 and X_2 are usually measured in different units, B_1 and B_2 are interpretable in terms of these original units but are not directly comparable. (That is, if $B_1 > B_2$, this does not necessarily indicate that X_1 is a stronger predictor of Y than X_2—it could be that a one-unit change on X_1 is much greater, as a proportion of the theoretical range of this variable, than a one-unit change on X_2.) To enhance comparability (especially when the units of X_1 and X_2 are arbitrary), investigators may choose to report the standardized regression coefficients, usually notated as β_1 and β_2. The standardized regression coefficient β_1 estimates the predicted change in Y, in standard deviation units, corresponding to a 1-*SD* change in X_1. Thus, if $\beta_1 = .25$, then a 1-*SD* change on X_1 is predicted to correspond to a 0.25-*SD* change on Y (holding X_2 constant).

Other effect sizes of importance in MRC include r (the bivariate correlation between two variables), R^2 (the squared multiple correlation coefficient, representing the proportion of Y variance accounted for by the predictor variables collectively), and sr^2 (the squared semipartial correlation, representing the proportion of Y variance uniquely accounted for by a single IV). The meaning of these different effect size indicators is discussed in Hoyt, Leierer, and Millington (2006).

Uses of MRC

Aiken and West (2000) described three broad uses of multiple regression analyses in psychological research:

> (1) *description,* to provide a statistical summary of the relationship of the *X*s to the *Y*; (2) *prediction,* to provide an equation that generates predicted scores on some future outcome (e.g., job performance) based on the observed *X*s; and (3) *explanation or theory testing.* In the third application, the sign and magnitude of predicted relationships of *X*s to *Y* can be tested using the actual observed data. (p. 350)

All three applications may be relevant to rehabilitation research, but we wish to focus special attention on the theory-testing function. We believe that this is the most common usage of MRC in the published literature and has the most direct implications for readers and reviewers evaluating the quality of research that uses these techniques. Indeed, Aiken and West (2000) asserted that it is this theory-testing application that "contributes most to the development of psychology as a basic science" (p. 350).

When researchers use MRC for theory testing, they are virtually always testing causal hypotheses. This goal of providing support for theories about causal relations among variables is a daunting one, especially when (as is often the case) the data being analyzed were collected at a single point in time and none of the variables has been experimentally manipulated. This reality encourages caution about the conclusions that can be drawn from study findings. (Editors may sometimes weary of having to remind researchers not to make causal inferences based on correlational data.) However, it does not change the fact that, at bottom, research findings are most useful if they inform theories—ideally, in an applied context, theories that in turn inform practice.

To put this another way, when researchers conduct regression analyses, they are usually building a primitive sort of causal model. For example, we might believe that X causes Y by virtue of X's intervening relationship with M (a mediator hypothesis). If this is true, then X, M, and Y should display a particular pattern of relationships in two regression equations. (See below for details.) If we observe this pattern in our sample data, we have not thereby proven that X causes M and, in turn, Y. We have shown that the observed data fit the predicted pattern, which provides some support for the theorized causal linkages. (To be rigorous about our inferences, we might conclude that the theorized linkages cannot be ruled out on the basis of our sample data.)

As in most causal modeling, the theoretical justification (which usually includes references to prior empirical findings) for the proposed linkages in MRC is at least as important as the effect-size estimates showing the strength of association between variables in our sample. To emphasize the central role of theory in causal modeling, Mueller (1997) cited two early advocates of the approach: "The study of structural equation models can be divided into two parts: the easy part and the hard part" (Duncan, 1975, p. 149). "The easy part is mathematical. The hard part is constructing causal models that are consistent with sound theory. In short, causal models are no better than the ideas that go into them" (Wolfle, 1985, p. 385).

This same point applies to MRC when it is used to test causal theories. Unless a strong argument (combining theory and empirical evidence) can be made for the proposed causal relations, interpretation of findings is ambiguous, and the potential of these findings to contribute to the cumulation of knowledge in the research area is critically compromised. Strengthening of ties between analyses and theories may be the single most important thing that researchers can do to enhance the scientific contribution of studies involving MRC.

Terminology

In this article, we use the term *independent variable* (IV) interchangeably with *predictor variable* and the term *dependent variable* (DV) interchangeably with *criterion variable*. This convention is not uncommon among textbook presentations of MRC (e.g., Cohen, Cohen, West, & Aiken, 2003; Hays, 1994; Pedhazur, 1982) and is consonant with the perspective just articulated, that research findings have the greatest scientific import when they are linked to causal theories about associations between variables. In this context, then, IV does not refer to a variable that is manipulated by the experimenter and therefore may be presumed on logical and empirical grounds to be the cause of associated changes in the DV (although this may sometimes be the case). Rather, it refers to a variable, usually measured rather than manipulated, that is presumed on theoretical grounds to have causal priority (i.e., to be a cause, rather than an effect, of other variables in the model). Lacking the strong warrant for causal inferences afforded by experimental control, researchers using observational methods are obliged to make a clear and compelling case for the hypothesized causal connections among constructs, to link the research design and analyses to these theory-derived hypotheses, and to consider explicitly (and rule out, when possible) competing explanations for their findings.

Model and Theory

One reason for linking research questions closely to theory is to increase the scientific yield of the study (Wampold, Davis, & Good, 1990). Rigorous tests of theory-derived hypotheses contribute to theory development, in that the conclusions of such a study will tend to strengthen confidence in valid theories or weaken confidence in questionable or incomplete theories. A second benefit of this linkage is the guidance it provides for research design and analysis. When the analytic model has been chosen to test a well-defined, theory-derived research hypothesis, the conceptual yield of the analysis is unambiguous, and interpretation of findings is straightforward. A flexible strategy for matching analysis to theory in MRC is the model comparison approach known as hierarchical regression analysis (HRA). We turn to this technique next.

Hierarchical Regression Analysis

A useful tool for researchers using MRC is the hierarchical regression strategy whereby the order of entry of IVs (or sets of IVs) into the regression model is predetermined to address questions of theoretical interest. HRA (which is also called *sequential regression* and is a special case of the model comparison approach advocated by Maxwell & Delaney, 2004) is really a series of regression analyses in which additional predictors are added at each step, to examine whether each new set of predictors accounts for significant variance in Y with the previously entered predictors still included in the model. Note that HRA should not be confused with hierarchical linear modeling (HLM; also known as multilevel modeling). In HRA, *hierarchical* refers to the prioritization of (sets of) IVs for entry into the regression equation; in HLM, it refers to a nested or hierarchical data structure that must be taken into account to avoid biasing results (Wampold & Serlin, 2000; see Kwok et al., 2008).

Application to predictive validity testing. Incremental validity analysis is a straightforward example of the use of HRA. When researchers assess the predictive validity of a new measure for use in an applied setting, it is often of interest to ask whether the new measure explains variance in the criterion over and above what can be accounted for by predictor variables already in use in the applied context. This incremental variance explained represents the value of adding the new measure to an existing predictive battery. Schmidt and Hunter's (1998) meta-analysis of predictors of job performance provides an illustration of the rationale for incremental validity studies. They examined the incremental validity of 18 categories of personnel measures as predictors of job performance over and above variance predicted by general mental ability (GMA). Schmidt and Hunter based their analyses on meta-analytically derived correlation matrices, to obtain stable estimates of the relevant validity coefficients. At Step 1 (sometimes called Block 1), they regressed job performance onto GMA, and at Step 2, they entered the second predictor type to see whether it explained significant additional variance in the DV. Several classes of measures (such as work sample tests, structured interviews, and integrity tests) were shown to improve predictive validity to a statistically significant degree compared with tests of GMA alone. Other potential predictors, such as job experience, years of education, and vocational interests, added little incremental validity. The critical factor in this analysis is the change in the variance explained when the new scale is added to the prediction equation.

Application to theory testing. HRA is also useful for hypothesis testing when hypotheses can be framed in terms of added or incremental Y variance accounted for by one set of predictors over and above what was explained by predictors entered at earlier steps in the model. For example, a researcher might be interested in the effect of religious engagement on mortality for a high-risk population, such as cancer patients. A significant association between religiosity and mortality at a fixed point in time (e.g., 5-year follow-up) would tend to support this hypothesis and could be tested via logistic regression, a variant of MRC used when the DV is categorical (e.g., alive or dead) rather than continuous.

Detractors of the hypothesis that religiosity affects physical health might argue that such a finding could be attributed to the effects of an intervening variable, such as social support, already known to predict health outcomes. Figure 1 depicts this alternative explanation for the observed association. This represents a mediational explanation for the observed effect. It does not challenge the supposition that religiosity is causally related to mortality but speaks to the mechanism that drives this relationship. Religiosity leads to an increase in social support, which in turn improves health, and there is no direct effect of religiosity on survival that might argue for the salutary effects of religiousness per se.

This explanation is particularly compelling when religiosity is measured by behaviors such as attendance at religious services (McCullough, Hoyt, Larson, Koenig, & Thoresen, 2000) because religious attendance (compared with other measures of religiosity such as religious attitudes or self-reported frequency of prayer) entails participation in a social activity, which is likely to lead to increased social support (Path *a* in Figure 1). The test for this hypothesis (assuming that a measure of social support is available in the data set) can be accomplished via HRA, with social support entered at Step 1, followed by religiosity at Step 2. The change in R^2 (or ΔR^2) at Step 2 then represents the incremental variance accounted for by religiosity over and above that explained by social support. If ΔR^2 is statistically significant, then social support cannot completely explain the bivariate association between religiosity and survival; if nonsignificant, then the effects of religiosity on health may best be attributed to the intervening role of social support.

When is HRA necessary? Astute readers may already have noted that HRA is not strictly needed in the example just described. When there are only two predictor variables (religiosity and social support), the DV may be simultaneously regressed onto both. The significance test for the regression coefficient for each IV assesses its unique relationship to the DV, while the other IV is statistically controlled. (In point of fact, the p value for this test will be identical to that for ΔR^2 in the HRA described above, and

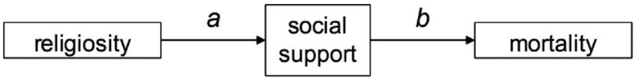

Figure 1. Social support accounts for the religiosity–mortality association.

the *t* statistic that tests whether *B* is different from zero, when squared, is identical to the *F* statistic that examines the same null hypothesis for ΔR^2). So, although HRA is a perfectly good strategy for addressing the hypothesis depicted in Figure 1, some might consider it to be overkill because the same conclusions can be derived from a simultaneous analysis.

HRA is advantageous when constructs are measured by sets of IVs, rather than individual IVs. For example, the study might include both a behavioral measure (attendance at religious services) and an attitude or affect measure (perceived level of connectedness to religious community) of religiosity. If both variables are entered as a set at Step 2, then ΔR^2 indicates their combined contribution to explaining variance in *Y,* whereas the regression coefficient (or sr^2, which is a measure of variance accounted for) estimates the unique contribution of each predictor in the set.

Another example of an IV that is naturally implemented as a set of variables in MRC is any predictor variable at the nominal level of measurement. Nominal (categorical) variables can be represented in HRA as a set of numerically coded variables. The number of code variables needed is equal to one less than the number of categories in the nominal variable. (See Cohen et al., 2003, chapter 8, for a detailed discussion of coding options for nominal variables in MRC.) So, if religiosity in the above study were measured as affiliation with one of four religious groups (e.g., Catholic, Protestant, Jewish, and other), HRA would again be the preferred approach to examining the contribution of the set of IVs that carry the religious group information to explaining variance in *Y.*

Challenges to applying and interpreting HRA. Petrocelli (2003) surveyed five volumes each of *Journal of Counseling Psychology* and *Journal of Counseling and Development* to see how researchers were using HRA in these journals. He found that about half of the articles that relied on regression procedures used HRA but that problematic practices were not uncommon. Problems identified included

> (a) lack of clarity as to whether the study was designed to explain or predict specific outcomes; (b) hypotheses that are not consistent with those that are testable with hierarchical regression; (c) lack of a clear, explicit rationale . . . ; (d) a focus on maximizing prediction rather than on theory-testing and the relative importance of predictor variables; (e) failure to examine and address probable problems of multicollinearity . . . ; (f) a discussion of results that focuses on the overall model rather than differences found through comparing progressive steps. (Petrocelli, 2003, p. 12)

Note that many of these problems reflect a lack of a clear, theory-driven rationale for conducting the HRA and consequent challenges to interpretation of findings—another reminder of the importance of grounding data analysis in psychological theory.

Petrocelli (2003) provided a number of instructive examples of studies that may not have gotten the most out of the HRA analyses conducted, although we note in passing that a number of studies were criticized for failing to enter causally prior variables at an earlier step than causally posterior variables. This is a practice we believe is not necessarily problematic. Indeed, in Figure 1, religiosity is assumed to cause social support, but it was necessary to enter social support at Step 1 to evaluate whether religiosity was uniquely related to *Y* even after accounting for its indirect effect on *Y* via social support. Thus, the "principle of causal priority" (which states that causally prior variables should be entered first; Petrocelli, 2003, p. 13) can be a useful rule of thumb, particularly when the goal of the HRA is to give credit to the *X* variables for the *Y* variance attributable to each (taking causal precedence into account). However, as we have seen, there are numerous applications of HRA, and some legitimately require variables later in the causal chain to be entered into the regression equation before the variables that cause them. The crucial consideration in determining whether an HRA is appropriate is whether the analysis has been conducted in such a way that the incremental variance explained (i.e., ΔR^2) provides a test of the hypothesis in question.

Use of HRA to statistically control for covariates is a common practice and is another example of the need for careful theoretical analysis. Jaccard, Guilamo-Ramos, Johansson, and Bouris (2006) provided an informative discussion of the pitfalls of *atheoretical partialing,* which refers to "the inclusion of covariates in a [regression] equation without careful consideration of their overall role in the broader theoretical network being tested" (p. 459). Their analysis showed that even seemingly innocuous procedures such as controlling for participant gender (by including gender among other covariates entered into the regression equation at Step 1) can result in biased parameter estimates or standard errors for the causal relation between the focal IV and the DV, depending on the nature of the causal relations among the covariate, the IV, and the DV.

For example, Nielsen (2003) studied the association between social support and posttraumatic stress disorder (PTSD). She entered a number of covariates at Step 1 and a pair of social support measures as a set at Step 2, looking to ΔR^2 (and its associated *F* test) as an indication that social support was still causally related to PTSD with the covariates statistically controlled. This procedure is such a common practice in psychological research that one could even characterize it as standard, but Jaccard et al. (2006) pointed out, following Meehl (1971), that automatic partialing of available covariates can have unsuspected consequences for validity of hypothesis testing studies. Statistical control for covariates is a valid use of HRA, but this step should always be considered in light of the causal relation of these covariates to the primary variables of interest and, thus, the implications of holding them constant for the interpretability of findings.

To illustrate, one of the covariates that attained statistical significance at Step 1 in Nielsen's (2003) analysis was marital status. Because marital status is related to social support (and in fact would generally be regarded as one important source of perceived social support), controlling for this variable very likely reduced the social support–PTSD association. The shared variance between marital status and the DV (as identified by the statistically significant regression coefficient for marital status) was partialed out of the DV at Step 1, leaving only the remaining *Y* variance to be explained by the social support measures at Step 2. There could be substantive reasons for this analytic strategy (if the focus of the research hypothesis were on extramarital support), but if overall support is the IV of interest, it seems likely that controlling for marital status leads to an underestimate of its association with PTSD. In general, Jaccard et al. (2006) suggested that "it is not sufficient to simply include all of the predictors and covariates into one large regression equation. Greater thought must go into the types of causal relations that may be operating" (p. 462).

Empirical (Stepwise) Regression: An Atheoretical Approach

Empirical or stepwise approaches to MRC can be attractive to researchers because they obviate the requirement we have been discussing of linking research to theory. In empirical regression, the data, rather than the researcher, make the choice about order of entry of the variables into the regression equation. In perhaps the most common (step-up) approach, the IV with the largest bivariate correlation with the DV enters first, followed by the IV that adds most to the variance explained, and so on, so that at each step, the new predictor that accounts for the greatest amount of incremental variance (over and above that explained by those already in the equation) is entered. This procedure concludes when none of the as yet unentered variables can make a significant incremental contribution to prediction and yields a streamlined predictor set that explains a large amount of *Y* variance.

Empirical regression methods have been relatively popular in *Rehabilitation Psychology*. In our survey of the issues from February 2004 to February 2008, we found 71 articles that used MRC analyses, and 14 of these (20%) used some form of empirical regression analysis. Hoyt, Leierer, and Millington (2006, p. 226) briefly summarized the statistical objections to this family of methods. We here note another objection to empirical regression, on conceptual grounds. These analyses inherently produce atheoretical findings that tell us nothing about the structure of associations among the variables in the equation. In considering these methods, Judd and McClelland (1989) commented that “it seems unwise to let an automatic algorithm determine the questions we do and do not ask about our data” (p. 465). As we have argued here, it is the challenging task of working out the causal relationships among variables that will advance scientific understanding of the phenomena under study. We believe that empirical regression methods should play a very limited role in scientific inquiry and are never an appropriate method for testing scientific theories.

MRC Approaches to Testing Mediator and Moderator Hypotheses

Baron and Kenny (1986) made a substantial contribution to the linkage of MRC analysis to theory in psychological research with their influential 1986 article distinguishing mediator relations from moderator relations and discussing how each type of hypothesis could be tested within an MRC or causal modeling framework. Recently, Frazier, Tix, and Barron (2004) revisited this topic, offering illustrations of the recommended methods applied to research questions of interest to applied psychologists and discussing newer developments and recommendations for investigators wishing to examine mediator and moderator relations in their data. We briefly review the basic techniques here and discuss controversies and refinements that have arisen in the past decade concerning both definitions and appropriate analyses of mediator and moderator hypotheses.

Tests of Mediation in MRC

When previous research has demonstrated an association between an IV and a DV, investigators may wish to examine proposed mediators of this presumed causal association. A mediator is an intervening variable caused by the IV, which in turn causes the DV, so that at least part of the effect of IV on DV is explained by its indirect effect via the mediator. The status of the mediator as an intermediate link in a causal chain can be illustrated by the example of a line of three dominos that are standing on end. When the first domino (the IV) is toppled, it will ultimately affect the final domino in the chain (the DV) but only because it first upsets the middle domino (the mediator), which in turn knocks over the final one.

Mediator hypotheses are important for theory development, and often have applied implications as well. For example, rehabilitation following traumatic injury generally involves a process of skill acquisition (or reacquisition), which includes a regimen of repetitive practice. Bandura (1977) theorized that practice has its effect on performance partly through the intervening variable of self-efficacy. That is, repetitive practice enhances self-efficacy, which in turn enhances performance. This mediator hypothesis, if supported, is important theoretically, for understanding the effects of practice and may have implications for rehabilitation counselors as well. For example, if self-efficacy is found to be a proximal cause of improved performance, then counselors may wish to employ other strategies for enhancing self-efficacy (in addition to repetitive practice). Vicarious learning, verbal reinforcement, and amelioration of performance anxiety are additional strategies for increasing self self-efficacy (Bandura, 1977) and may be instrumental in facilitating skill acquisition.

Numerous analytic strategies have been developed for testing mediator hypotheses (see MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; Shrout & Bolger, 2002), but the simplest and by far most popular involves a straightforward application of MRC proposed by Baron and Kenny (1986). This procedure involves three steps:

1. Regress the mediator onto the IV, to show that it is plausible that these two variables can be causally linked.
2. Regress the DV onto the IV, to show that a causal relation is plausible here as well.
3. Regress the DV simultaneously onto the IV and the mediator, to show that the mediator is significantly related to the DV even when the IV is statistically controlled.

If the regression coefficients for Steps 1 and 2 are significant and the partial regression coefficient for predicting the DV from the mediator is significant in Step 3, then a mediator hypothesis is supported (Baron & Kenny, 1986).

The path models depicted in Figure 2 are helpful for understanding this procedure. In these models, *X* represents the IV, *Y* represents the DV, and *M* represents the mediator. Path *a* is the effect of *X* on *M* (tested in Step 1 above). Path *b* is the effect of *M* on *Y,* controlling for *X* (tested in Step 3 above). Both of these paths must be significant to support a mediator relation between *X, Y,* and *M.* In addition, Baron and Kenny’s (1986) formulation requires that Path *c* (the effect of *X* on *Y,* ignoring *M*) be statistically significant (Step 2 above). That is, it does not make sense to test for mediators of the *XY* relation unless this relation is statistically significant. More recently, Kenny, Kashy, and Bolger (1998) argued that this requirement may be relaxed in some circumstances. For example, when two different mediating variables produce contrasting effects

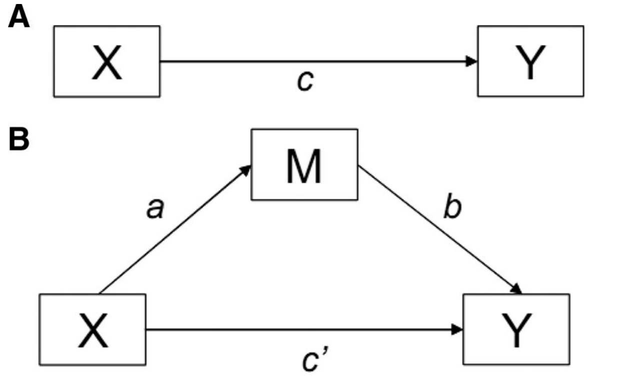

Figure 2. A: Bivariate association of *X* and *Y*. B: *M* as mediator of the effect of *X* on *Y*.

on the DV, it may be important to examine mediator hypotheses even when the bivariate correlation between *X* and *Y* is near zero or nonsignificant (MacKinnon et al., 2002). When a long time interval elapses between the measurement of *X* and *Y,* their bivariate correlation may be expected to be relatively weak, yet mediator hypotheses may still be of theoretical and practical interest (Shrout & Bolger, 2002).

Example mediator analysis. It is documented that chronic pain patients have a high propensity to catastrophize (Geisser, Robinson, & Riley, 1999; Jones, Rollman, White, Hill, & Brooke, 2003; Turner, Mancl, & Aaron, 2004). In the context of chronic pain, catastrophizing is a cognitive process characterized by a lack of confidence and control and an expectation of negative outcomes (Keefe, Brown, Wallston, & Caldewell, 1989) and may be implicated in the development of depression in the afflicted patients. Lewinsohn, Hoberman, Teri, and Hautzinger's (1985) integrative model of depression suggests that disruption of scripted behaviors in daily routines (e.g., due to a work injury) will reduce positive reinforcement. When coupled with activity interference because of the injury (further leading to reductions in positive reinforcement), this disruption may lead to cognitive distortions (e.g., catastrophizing) and, in turn, to depression. Researchers have demonstrated a positive association between catastrophizing and higher levels of disability, higher rates of health care usage, longer hospitalizations, increased pain medication usage, and longer time to reach rehabilitation milestones (Banks & Kerns, 1996; Jones et al., 2003; Keefe, Rumble, Scipio, Giordano, & Perri, 2004; Turk, 1999, 2003; Turk & Okifuji, 2002; Turner et al., 2004).

We now test a mediational hypothesis of the relation between pain intensity and catastrophizing in a sample of 171 persons receiving workers' compensation from Lee, Chan, and Berven's (2007) study of predictors of depression in individuals with chronic musculoskeletal pain. Lee et al. recruited participants from six outpatient rehabilitation facilities in the province of Alberta, Canada. To be included in Lee et al.'s study, persons had to be 21 years or older with a medical diagnosis of nonmalignant, work-related musculoskeletal pain for at least 3 months based on the criteria of chronicity specified by the International Association for the Study of Pain (IASP) Subcommittee on Taxonomy (IASP, 1986). Correlations and descriptive statistics (for the 141 cases with complete data on the variables used in this and other illustrative analyses in this article) are provided in Table 1.

Following Lewinsohn et al.'s (1985) theory, we hypothesized that the association between pain intensity and catastrophizing is mediated by both stress, which may contribute to irrational thought processes, and activity interference, which will lead to too much time for rumination and create feelings of hopelessness and futility, fueling cognitive tendencies to expect the worst. This model is depicted in Figure 3 and was tested using four separate regression equations. As expected, the association between pain intensity and catastrophizing (Step 2 above) was significant: β (95% confidence interval [CI]) = .46 (.32, .58). (Note that when the 95% CI excludes 0, the effect size differs significantly from zero, $p < .05$.)

Next, we assessed the association between the IV and each of the mediators (Step 1 in the Baron & Kenny, 1986, analysis). Pain intensity was significantly related to both actual stress and activity interference, βs = .24 (.08, .39) and .45 (.31, .57). Finally, we examined whether each of the putative mediators was significantly related to the DV, while statistically controlling for the IV (Step 3). The relevant analysis was a simultaneous regression of catastrophizing (DV) onto pain intensity (IV), actual stress, and activity interference (both mediators). (The reason for including both mediators and the IV in the regression equation is that the path model in Figure 3 implies that each mediator is uniquely related to catastrophizing, controlling for the other mediator and for the IV.) In this analysis, both mediators were significantly associated with the DV, βs = .34 (.17, .50) and .24 (.09, .38) for actual stress and activity interference, respectively.

In summary, all three steps were significant as predicted, yielding support for the proposed dual-mediation model. That is, the findings conform to the predictions of a model in which pain is related to catastrophizing indirectly through its association with both stress and activity interference, each of which is uniquely related to catastrophizing. Finally, we examined the association between pain and catastrophizing in the final (three-predictor)

Table 1
Correlations (and 95% Confidence Intervals), Ms, and SDs for Variables Used in Example Analyses (N = 141)

Variable	1	2	3	4	5
1. Pain	—				
2. Catastrophizing	.41 (.27, .54)	—			
3. Total stress	.27 (.11, .42)	.44 (.29, .56)	—		
4. Social support	−.12 (−.28, .05)	−.39 (−.52, −.24)	−.49 (−.61, −.36)	—	
5. Activity interference	.45 (.30, .57)	.52 (.39, .64)	.44 (.30, .57)	−.38 (−.52, −.23)	—
M	58.48	13.42	707.81	43.03	4.26
SD	27.02	7.82	361.71	11.96	1.14

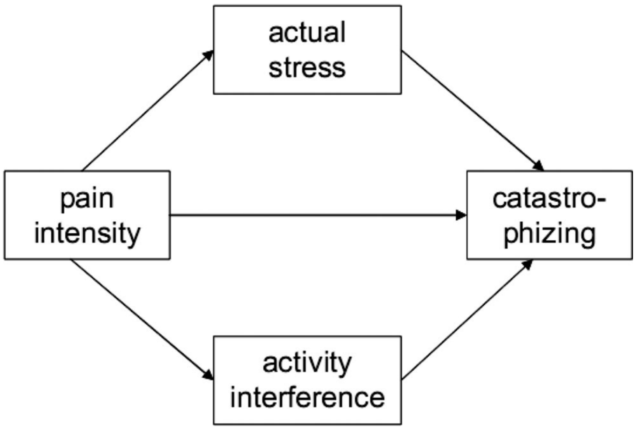

Figure 3. Dual-mediation model of the effect of pain intensity on propensity to catastrophize.

regression equation. This was also statistically significant, β = .20 (.04, .36). Thus, while the intervening variables (stress and activity interference) help to explain the link between pain and catastrophizing, they do not completely explain it. This pattern of findings, with significant indirect effects through one or more mediators accompanied by significant direct effects, is not uncommon and was characterized by Baron and Kenny (1986) as a case of partial mediation.

Test of indirect effect. A complement to Baron and Kenny's (1986) approach to testing mediator hypotheses is the test of the significance of the indirect effect of the IV on the DV through the mediator. Numerically, this indirect effect is equal to the product of the two path coefficients that compose the indirect path from *X* to *Y* in Figure 2B (i.e., *ab*). A simple approach to testing whether this product differs significantly from zero was proposed by Sobel (1982). Sobel assumed a normal sampling distribution for *ab* and used an approximate standard error for this product to create a CI (using a standard procedure described later in this article) or conduct a significance test.

Unfortunately, the sampling distribution of the product *ab* does not usually approximate a normal distribution, so that significance tests using Sobel's (1982) method are biased (MacKinnon et al., 2002; Mallinckrodt, Abraham, Wei, & Russell, 2006; Preacher & Hayes, 2004). Several alternative tests have been proposed (MacKinnon et al., 2002), and a consensus appears to be emerging that the test with the best statistical properties is the bootstrap (Shrout & Bolger, 2002). Mallinckrodt et al. (2006) presented procedures for conducting bootstrap tests of indirect effects for many common statistical software packages, and a user-friendly macro for conducting bootstrap tests of single-mediator and multiple-mediator models can be downloaded (for either SPSS or SAS) from Kristopher Preacher's Web site (www.people.ku.edu/~preacher/). For our dual-mediation model (Figure 3), the unstandardized indirect effects (with 95% CIs derived from bias-corrected and accelerated bootstrap procedures) were $ab = 0.020$ (0.006, 0.041) and 0.047 (0.024, 0.077) for the indirect paths through stress and activity interference, respectively. As expected, given that the individual paths *a* and *b* were significant for each mediator, the 95% CIs do not include zero, indicating that the indirect effect is significant ($p < .05$) in each case.

Cautionary note regarding mediator analyses. An important caveat pertains to mediator analyses using IVs that are measured rather than experimentally manipulated. In such studies, the direction of causality is presumed, rather than empirically confirmed, and it is crucial to provide a strong theoretical justification for the effects depicted in Figure 2. Researchers studying mediator associations between measured variables can strengthen the basis for causal inference by collecting longitudinal data. In fact, the MacArthur Group (Kraemer, Wilson, Fairburn, & Agras, 2002) recommended that temporal precedence (i.e., measurement of the IV at an earlier time point than the mediator) should be an additional criterion for tests claiming to provide evidence of mediation in intervention research.

Although the temporal precedence requirement may be relatively easy to meet in clinical trials (where the IV is assignment to the treatment or comparison group), it is often not met in theory-testing investigations that treat measured variables as IVs. Maxwell and Cole (2007) surveyed five American Psychological Association (APA) journals for studies testing mediator hypotheses and found that 38 of the 72 separate studies they identified tested these hypotheses in cross-sectional data (i.e., the IV, mediator, and DV were all measured at the same point in time). This was also the case with our example study, which creates challenges for drawing causal inferences. In particular, it is essential for researchers conducting mediator tests on cross-sectional data to recognize that no empirical support is offered by such analyses for the presumed direction of causation. Because the hypothesized causal linkages are theoretically rather than empirically justified, researchers should feel obliged to explicitly consider alternative causal models and to evaluate their plausibility on theoretical grounds.

In our example analyses, an alternative explanation for the observed associations between variables is that catastrophizing causes pain intensity, rather than the other way around. Figure 4 depicts a possible alternative model presuming that distorted thought processes, characterized by a propensity for rumination and a pessimistic mindset, can accentuate perceptions of pain intensity. The causal relation postulated between stress and catastrophizing is consistent with Beck's (1967) cognitive theory of depression. Beck postulated that individuals develop patterns of distorted thinking early in life that create vulnerabilities to depression and that these cognitive schemas are likely to become activated in periods of high stress. Thus, Figure 4 depicts an alternative mediational model in which catastrophizing mediates the association between stress and pain intensity. This model was also

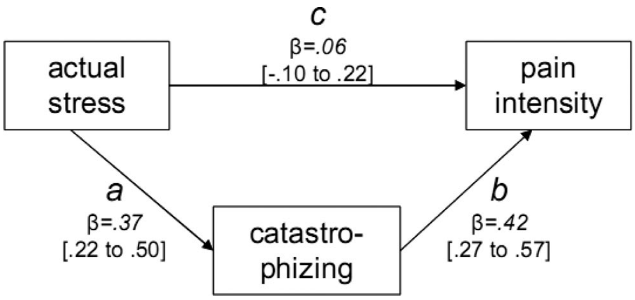

Figure 4. Alternative mediational model (with standardized path coefficients and 95% confidence intervals) for pain data.

well supported by the data and in fact provides an example of full mediation, in that stress no longer has a significant association with pain intensity when catastrophizing is statistically controlled.

The lesson to be learned here is that, similar to more elaborate causal models based on variables measured at a single time point (MacCallum, Wegener, Uchino, & Fabrigar, 1993), interpretation of the results of cross-sectional mediation tests is perilous. Investigators who fail to provide strong theoretical and empirical justification for the proposed structural relations among variables may well be testing implausible models of mediation and very likely are ignoring equally plausible alternative models. The solution to this problem is to strengthen the basis for causal inferences by measuring the IV and the mediator at different time points (Kraemer et al., 2002) and ideally also measuring the mediator and the DV at different time points (Maxwell & Cole, 2007). When this is not possible, it is critical that investigators give due consideration to alternative models and show why these are less plausible, on theoretical grounds, than the model under investigation.

Is Baron and Kenny's (1986) Approach Outmoded?

In a Monte Carlo study comparing 14 methods for testing mediation, MacKinnon et al. (2002) concluded that Baron and Kenny's (1986) approach has Type I error rates that are too low and "very low power" (p. 96) relative to more recently developed mediation tests. The implication seems to be that investigators wishing to test mediator hypotheses should eschew this well-known method and replace it with tests of the significance of the indirect effect *ab* we described earlier (e.g., MacKinnon et al., 2002; Shrout & Bolger, 2002). This conclusion appears to have been taken to heart in some segments of the applied psychology research community (e.g., Frazier et al., 2004; Mallinckrodt et al., 2006). Nonetheless, we advocate the use of the Baron and Kenny method, with one possible modification, as a default test of mediation hypotheses. The modified Baron and Kenny approach has excellent statistical properties, provides more information than alternative approaches testing a combined indirect effect, embodies a sound theory of mediation, and is readily conducted using existing statistical software and easily interpreted by readers familiar with MRC. Because this recommendation appears to be controversial, we elaborate on our reasons for this preference.

Statistical properties. The poor power of the Baron and Kenny (1986) method in MacKinnon et al.'s (2002) simulations is exclusively a function of the fact that this approach requires that the *XY* relation be significant (Step 2, above), whereas the alternative mediation tests do not. The difference in power was particularly acute in MacKinnon et al.'s published tables, data for which were generated from populations with full mediation—that is, populations in which the direct relation between *X* and *Y,* controlling for *M* (path c' in Figure 2B) is zero. Examples of full mediation are probably relatively rare in the behavioral sciences (Baron & Kenny, 1986), and in such cases, it is not unusual for Paths *a* and *b* to be significant, whereas Path *c* is not. According to Baron and Kenny's original formulation, when Path *c* is not significant, there can be no point in testing for mediation—a guideline that in the MacKinnon et al. (2002) simulations produced large numbers of Type II errors for the Baron and Kenny approach, particularly for small mediation effects.

As noted above, recent treatments of mediation have made the case that, in at least some circumstances, it is appropriate to test for mediation in the absence of a significant *XY* association. Kenny et al. (1998) acknowledged that the requirement embodied in Step 2 of the original formulation may sometimes be relaxed. Relaxing this requirement would be expected to greatly reduce the rate of Type II errors. Indeed, MacKinnon et al. (2002) included in their simulations a test called *joint significance* of Paths *a* and *b*, which consists of Steps 1 and 3 of the Baron and Kenny (1986) approach, with the Step 2 requirement omitted. This test of joint significance had Type II error rates as low as or lower than all but the two most powerful of the newer mediation tests. Thus, when the Step 2 requirement is relaxed, Baron and Kenny's approach performs favorably relative to newer, more statistically sophisticated mediation tests. In addition, the two tests more powerful than this joint significance approach appear to be overpowered, yielding very high Type I error rates when either Path *a* or Path *b* is zero in the population (see MacKinnon et al., 2002, Table 9). These findings prompted MacKinnon et al. to state that "the best balance of Type I error and statistical power . . . is the test of joint significance" (MacKinnon et al., 2002, p. 83).

Definition of mediation. MacKinnon et al. (2002) distinguished three classes of mediator tests: (a) *causal steps tests* like that proposed by Baron and Kenny (1986); (b) *difference in coefficients tests,* which evaluate the significance of the change in the *XY* path with and without the mediator included in the regression model (i.e., $c - c'$ in Figure 2); and (c) *product of coefficients tests,* which evaluate the significance of a product term formed by multiplying the coefficients of the two (or more) paths that make up the indirect effect (i.e., *ab* in Figure 2). MacKinnon et al. favored the last two approaches because each defines a single effect size for the indirect effect and tests the significance of this effect size. We prefer the causal steps approach, as it conforms to the conception of mediation as links in a chain of causation and tests each of the links individually for significance.

Additional information. Testing and computing an effect size for each link in the causal chain yields additional information that may have implications for both theory and practice. Imagine that *X* in Figure 2 represents a job skills intervention, *Y* represents a vocational outcome such as hours of employment at 6-month follow-up, and *M* represents a behavioral measure of interpersonal skills, which is one focus of the training and is expected to enhance employability. A finding that the indirect effect (from *X* to *Y* through *M*) is significant and of moderate magnitude (e.g., $ab = .1$) would be encouraging but would have limited implications for improving the intervention. More useful is the information that $a = .2$ (and is statistically significant) and $b = .5$ (and is statistically significant). This tells us that the mediator is important (i.e., strongly related to the DV) but that the intervention is only modestly effective in creating changes in *M.* Contrast this pattern of associations with an alternative set of findings where $a = .5$ and $b = .2$. Here, the intervention is strongly associated with changes in *M,* but these changes are only modestly associated with employability. These two hypothetical situations are identical from the point of view of the product of coefficients test (and also of the difference in coefficients tests) but differ in their implications for understanding and improving intervention effectiveness—a difference that is highlighted when the causal steps test is used.

Straightforward test. In contrast to the difference tests and the product tests discussed by MacKinnon et al. (2002) and also to the bootstrap approaches presented by Shrout and Bolger (2002; Mallinckrodt et al., 2006), the modified Baron and Kenny (1986) test for mediation requires no specialized software and no statistical knowledge beyond the basics of MRC. As noted by Wilkinson and the Task Force on Statistical Inference (1999), researchers should make a practice of using the simplest method that is adequate to the nature of the research question. Researchers who use a method they understand well are more likely to understand and explain their findings effectively and more likely to notice irregularities that could be caused by outliers, data entry problems, or erroneous use of computer software, as compared with those who use methods that are outside their statistical comfort zone. In addition, the results of the simpler analysis will be more readily understood by readers.

In summary, Baron and Kenny's (1986) approach to testing mediator hypotheses (with Step 2 omitted, as appropriate) performs well relative to more complex testing procedures that directly test the significance of the indirect (i.e., mediator) effect. This method embodies an intuitive understanding of mediation as a chain of causation and provides effect sizes and precision estimates (*p* values or, preferably, CIs) for each link in the chain. In addition, Baron and Kenny's tests can be conducted using familiar statistical software, and their results are readily intelligible to both readers and researchers. This is a desirable characteristic: Wilkinson and the Task Force on Statistical Inference (1999) noted that "Occam's razor applies to methods as well as theories" and adjured their readers to choose a "minimally sufficient analysis" (p. 598) to extract relevant findings from their data. We believe that MRC-based approaches conform well to this criterion.

Finally, although it has been claimed that "it is difficult to extend the causal steps method to models incorporating multiple intervening variables" (MacKinnon et al., 2002, p. 87), such an extension is straightforward. The core of the causal steps method is that each intervening path in the causal chain must be statistically significant for a mediator hypothesis to be supported. The modified Baron and Kenny (1986) method uses focused tests in two separate regression analyses to examine the two links (Paths *a* and *b* in Figure 2) in a simple mediational chain, but this same logic applies to tests of a single chain with more than one mediator variable or to multiple chains each with one or more intervening variables. Jaccard et al. (2006) described a method they called *directed regression* for designing a series of regression analyses to provide tests of complex mediational relationships, and Kenny (1979) showed how to conduct more complex path analyses via multiple regression.

Testing Moderator (Statistical Interaction) Hypotheses

A moderator variable is defined as a third variable that affects the strength and/or direction of association between an IV and a DV (Baron & Kenny, 1986). Thus, mediation and moderation embody distinct roles for a third variable in explicating the association between two primary variables of interest. Mediator hypotheses (discussed above) investigate how the IV affects the DV; moderator hypotheses investigate when (i.e., under what conditions) or for whom this association is relatively strong or weak.

Figure 5 depicts a common shorthand for diagramming moderator relations, with a causal arrow from *X* to *Y*. The moderator variable (*M*) is not necessarily related to either *X* or *Y* but is represented as having its effect on the causal arrow that connects them. This arrows-pointing-at-arrows representation is not meaningful mathematically—Figure 5 is not a path diagram, although it certainly resembles one. It does, however, provide an effective visual representation of what it means for *M* to be a moderator of the *X*–*Y* association: The effect of *X* on *Y* is a function of *M*.

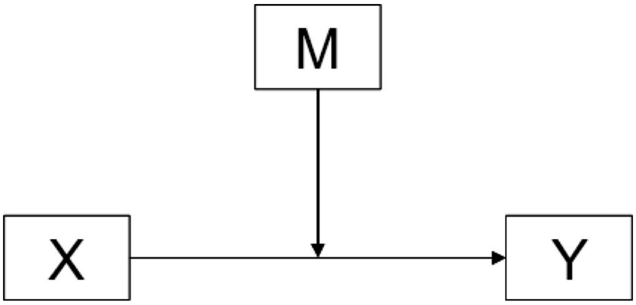

Figure 5. Conceptual diagram of a moderator hypothesis (*M* moderates the association between *X* and *Y*).

Figure 6 shows a second visual representation of the moderator test. This path diagram is mathematically accurate and clarifies the procedure for testing the moderator hypothesis, although it may be less helpful as a visual aid for conceptualizing the nature of the moderator relation. A moderator relation is also called a *statistical interaction* between two predictor variables with respect to a given DV. If the moderator hypothesis is correct, the two variables have a multiplicative, rather than a simple additive, relation to the DV. For this reason, testing a moderator hypothesis requires that the investigator create a new variable that is the product of the IV and the moderator variable (i.e., $X \times M$ in Figure 6). We illustrate this procedure with a test of a moderator hypothesis from our example data set.

Example moderator test. In our original mediator analysis, we tested Lewinsohn et al.'s (1985) model in which pain causes distorted thinking because it reduces participation in activities that result in positive reinforcement and increases stress. For the moderator analysis, we examine whether social support (*M* in Figures 5 and 6) moderates the association between pain (*X*) and catastrophizing (*Y*). Social support has been shown to be a buffer against many life stressors, and factors posited to have a buffering or protective role are often best conceptualized as moderators of the association between IV and DV. Note that in stating a moderator hypothesis, it is important to be explicit about the nature of the hypothesized moderator effect. Because social support is theorized to buffer the pain–catastrophizing relationship, we hypothesized that this IV–DV association will be weaker for persons who report high levels of support, as compared with those who report low levels of support.

Aiken and West (1991) recommended several steps to enhance interpretability of moderator findings. Prior to analysis, it is a good idea to center *X* and *M*. This reduces the collinearity of each of these predictor variables with the product term and also gives the regression coefficients greater practical meaning. (When both predictors are centered at their means, the intercept is the predicted value of *Y* for persons scoring at the mean on both *X* and *M,* and the B_X represents the slope of the *Y*-on-*X* regression line for persons scoring at the mean on *M*.) Centering is accomplished by

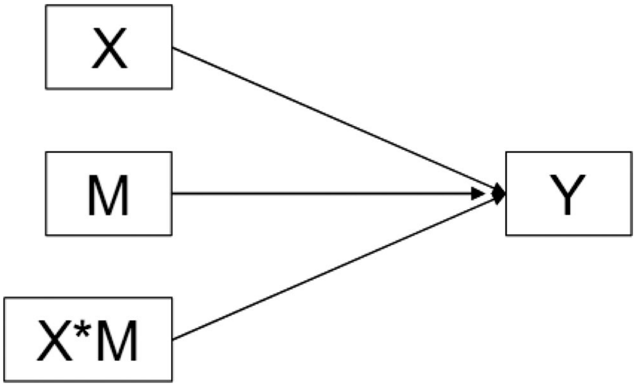

Figure 6. Mathematical depiction of the test of a moderator hypothesis (M moderates the association between X and Y).

computing the mean for each variable, which, in our case, yielded the values in Table 1, then creating two new variables (i.e., $Xcent = X - \bar{X}$, and $Mcent = M - \bar{M}$). These centered scores are then multiplied together to create a third predictor variable (i.e., $XM = Xcent \times Mcent$).

Once the three predictor variables have been created, the second step is to regress Y onto the three predictors (*Xcent, Mcent,* and *XM*). It is common to do this using HRA, as illustrated in Table 2, with the two main effects (X and M) entered at Step 1, and the product term added at Step 2. As stated in Table 2's note, $\Delta R^2 = .04$ at this second step in the regression, $F(1, 137) = 7.28$, $p < .05$. Because the product term contributes significantly to explaining variance in Y, there is evidence of moderation. That is, the relation of pain to catastrophizing differs depending on available social support.

As discussed earlier, this hierarchical moderator analysis is not strictly necessary when both X and M are represented in the regression equation by a single variable. In this case, only a single product term enters the equation at Step 2, and the t test for the significance of B in a simultaneous regression will provide the same information as the F test for the significance of ΔR^2 in HRA. (It is critical, however, to include both X and M in the equation when testing the significance of the product term XM. The XM association with Y tests for a statistical interaction only after variance accounted for by X and M individually has been partialed from Y.) If either X or M is represented by a set of two or more variables, then the interaction effect will necessarily be represented by two or more product terms, and the HRA approach becomes obligatory to obtain an omnibus test of the interaction.

Graphing interactions. When there is evidence of a moderator effect, a third important step is to graph the interaction, so that readers can understand the nature of the moderator relationship. Aiken and West (1991) recommended a graph of three simple regression lines, looking at the regression of Y on X at three different levels of M. The recommended levels are one standard deviation below the mean of M, the mean of M, and one standard deviation above the mean of M. These values cover a sizable proportion of the distribution of the moderator variable without being so extreme as to be atypical.

Figure 7 is a graph of the interaction between pain and social support in predicting catastrophizing. It contains the three simple regression lines recommended by Aiken and West (1991). If the interaction between the predictors is negligible, the three lines will appear parallel. The definition of a significant interaction is that the slopes of these regression lines differ (because the slope of each regression line represents β_X, the conditional effect of X on Y at a given level of M). The graph shows that the interaction between pain and social support in this sample is in the opposite of the predicted direction. When social support is low (top line in Figure 7), there is a relatively weak relation between pain and catastrophizing, and the slope of the conditional regression line (Y-on-X, at a given level of M) increases as social support increases. In other words, the effect of pain on catastrophizing becomes stronger as social support increases.

Figure 7 makes clear the nature of the multiplicative (i.e., moderator) relation between X and M as predictors of Y. Catastrophizing (Y) is minimized when pain (X) is low and social support (M) is high. But at low levels of social support, catastrophizing is likely to be relatively high regardless of pain intensity.

A second approach to probing significant moderator effects was introduced by Bauer and Curran (2005). They called it the *regions of significanc*e approach, or the *J-N technique* (because it is an adaptation of the technique introduced by Johnson & Neyman, 1936, for probing interactions between a categorical predictor and a continuous moderator in analysis of covariance [ANCOVA]). The goal of the J-N technique is to specify where in the range of the moderator variable the conditional effect of X on Y is statistically significant.

Note that Figure 7 tells us how the slope of the conditional Y-on-X regression line changes with changes in M, but it does not tell us which of these simple slopes differs significantly from zero. Bauer and Curran (2005) recommended computing a 95% CI for the simple slope parameter at each level of M. As always, when the 95% CI includes zero, the effect size (in this case, β_X) does not differ significantly from zero. By graphing the simple slope, along with its upper and lower confidence limits, as a function of the moderator (M), we can see where in the range of M the conditional Y-on-X regression is significant and where it is not.

Figure 8 is a graphical representation, using the J-N technique, of the regions of significance for the conditional regression of catastrophizing on pain at different levels of social support. The dark (dashed) line in this figure represents the predicted value of

Table 2
Social Support as a Moderator of the Pain-Catastrophizing Association (N = 141)

Variable	B	95% CI lower	95% CI upper	β
Step 1				
Pain	.13	.08	.17	.39*
Social support	−.22	−.31	−.13	−.33*
Step 2				
Pain	.128	.08	.17	.40*
Social support	−.24	−.33	−.15	−.37*
Pain × social support	.005	.001	.009	.19*

Note. $R^2 = .29$ for Step 1 (95% CI = .16, .41); $\Delta R^2 = .04$ at Step 2, $F(1, 137) = 7.28$, $p < .01$. Pain and social support scores were mean-centered prior to analysis. CI = confidence interval.
* $p < .05$.

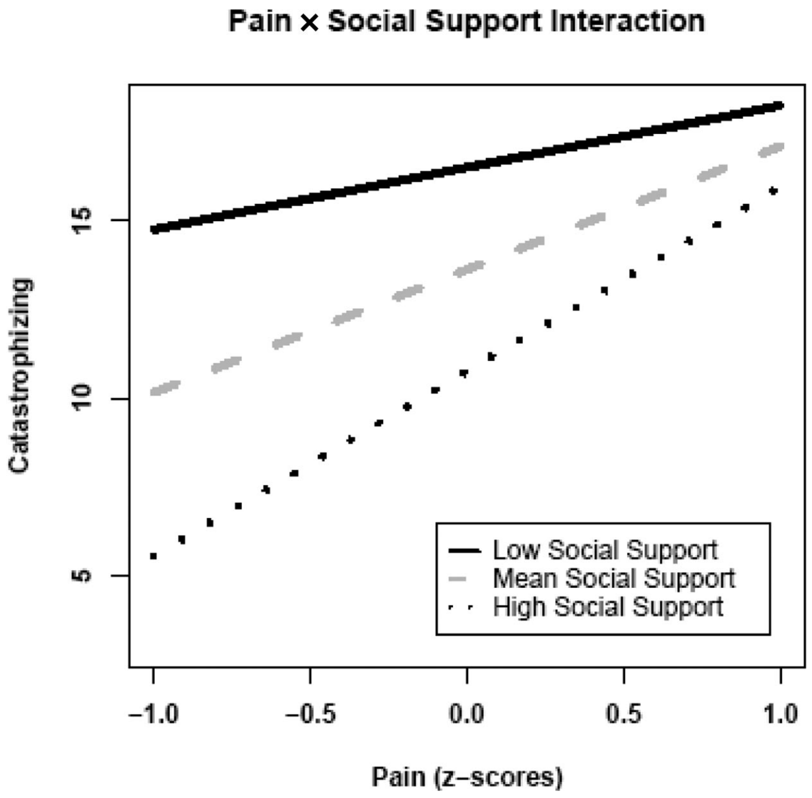

Figure 7. Conditional regression of catastrophizing on pain at three levels ($z = -1$, 0, and 1) of social support.

β_X (the slope of the corresponding simple regression line in Figure 7) at each level of the moderator variable. It is bracketed by a confidence band, and the regions of significance (delineated by vertical dashed lines) indicate where in the range of the moderator variable the simple slope is (a) significantly negative, (b) not significantly different from zero, and (c) significantly positive.

For routine research reports, it may not be necessary to publish a figure illustrating regions of significance for the conditional regression. It may be enough to report the critical values (z scores on the moderator variable in Figure 8) where the confidence limits cross the x-axis. In Figure 8, $\text{CL}_{\text{lower}} = 0$ when $z_{\text{SS}}= -1.01$, and $\text{CL}_{\text{upper}} = 0$ when $z_{\text{SS}}= -7.56$. That is, pain is predicted to be significantly (and positively) related to catastrophizing for persons 1.01 *SD* below the mean of social support, and this association grows stronger (and remains statistically significant) at higher levels of social support. Relating this information to Figure 7, each of the conditional regression lines depicted (at values of $z_{\text{SS}} = -1$, 0, and 1) depicts a significant conditional association between pain and catastrophizing. For persons scoring very low ($z_{\text{SS}} < -1.01$) on the moderator variable, however, the conditional association does not differ from zero, and the level of catastrophizing is independent of pain intensity. In theory, at extremely low levels of social support, there would be a negative relation between pain and catastrophizing. However, Figure 8 makes clear that this lower region of significance occurs at such extreme levels of the moderator variable ($z_{\text{SS}}= -7.56$) that it has little practical import.

Kristopher Preacher's Web site (www.people.ku.edu/~preacher/) includes a section on graphical analysis of interactions in MRC, with Java applets that assist researchers in probing interaction effects. Figures 7 and 8 were produced using these tools, and we recommend them for researchers seeking to convey the meaning of significant moderator effects in their own research.

Why not use a median split? Analysis of moderator effects with continuous predictor variables may be daunting initially, and investigators sometimes take what seems to be the expedient route of converting continuous variables into categorical variables, so that a more familiar analysis of variance (ANOVA) can be used to test the moderator hypothesis. For example, one could split both X and M at the median of the scale, assigning half of the sample to a high-pain group and the other half to a low-pain group (and similarly for social support). This creates a familiar 2 × 2 ANOVA design, in which main effects and interactions are readily assessed.

Although common, the practice of dichotomizing continuous predictor variables is highly problematic and should be avoided. Dichotomization of continuous variables for any type of analysis (not just moderator analyses) compromises statistical power and can yield misleading results (MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993). So, researchers studying naturally continuous variables are well advised to familiarize themselves with regression approaches to testing moderator hypotheses. Frazier et al. (2004) provided a detailed primer on this approach, and Aiken and West's (1991) book on the subject is unsurpassed for its clarity and comprehensiveness.

Statistical power of moderator tests. Aiken and West (1991, chapter 8) discussed the statistical power of moderator analyses. We summarize the most important points here and refer readers to

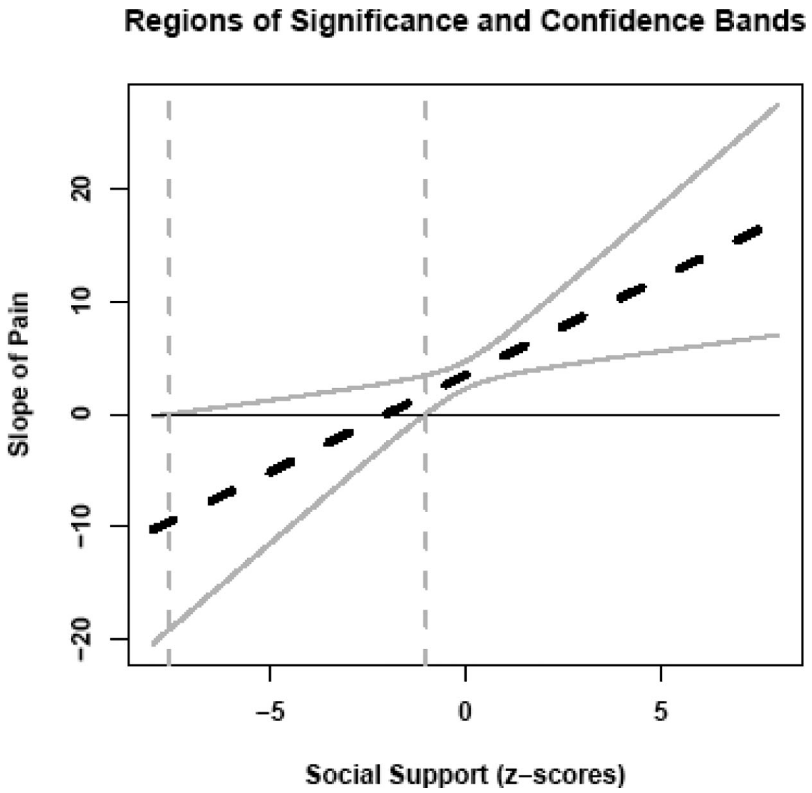

Figure 8. Conditional regression of catastrophizing on pain as a function of social support, with confidence bands and critical values delineating regions of significance for the simple slope coefficient.

the original source for more detailed evidence supporting these conclusions. Statistical power (i.e., the probability of correctly rejecting a false null hypothesis) in MRC is, as usual, a function of sample size, effect size, and alpha level (Type I error rate). The effect size that is directly related to statistical power is not ΔR^2 but, rather, the signal-to-noise ratio f^2, which may be written as the ratio $\Delta R^2/(1 - R^2_{\text{tot}})$, where R^2_{tot} is the squared multiple correlation after the product term enters the regression equation at Step 2. Aiken and West made two points about f^2 as a measure of effect size for moderator tests. First, Cohen's (1988) rules of thumb concerning small, medium, and large effect sizes ($f^2 = .02$, .15, and .35, respectively), whatever one may think of their suitability for HRAs in general, appear to overestimate the strength of moderator effects in psychological research. It is very unusual to observe a moderator effect as large as $f^2 = .15$, so this might well be considered a large effect size, as moderator effects go. Aiken and West recommended that an effect half that large (i.e., $f^2 = .075$) may be a reasonable estimate of a medium moderator effect, while $f^2 = .02$ may be reasonable as a benchmark for a small effect. Researchers who conduct power analysis to determine the appropriate sample size for moderator tests will likely derive an inappropriately small recommended sample size if they use Cohen's large $f^2 = .35$ (or even his medium $f^2 = .15$) as an estimate of the expected magnitude of the interaction effect.

A second implication of the formula for f^2 is that, for a given proportion of variance accounted for by the product term (ΔR^2 when *XM* enters the equation at Step 2), power will vary as a function of the additional variance explained at Step 1, which also contributes to R^2_{tot}. This helps to explain why so-called crossed (or disordinal) interactions are so rarely detected in psychological research. In a crossed interaction, the direction of association between IV and DV changes from high to low levels of the moderator variable. For example, the IV might be positively related to the DV at high levels of *M* but negatively related for participants with lower *M* scores. In a disordinal interaction, the main effect of *X* at Step 1 will be small because the strong effects for high-*M* and low-*M* participants are in opposite directions and tend to cancel one another out. This means relatively little variance accounted for at Step 1, which means a larger denominator for f^2 and a smaller overall effect size relative to an ordinal interaction that accounts for the same amount of variance (ΔR^2) at Step 2.

The final point that Aiken and West (1991) made about statistical power in moderator tests is that measurement error in the IV and the moderator combine to create a greater amount of error variance in the product term (which carries the interaction variance) than in either *X* or *M*. This compounding of error variance is partly a function of the correlation between *X* and *M*, but it is reasonable to expect that the proportion of error variance in the product term will be roughly double the proportion of error variance in *X* or *M*. Because measurement error acts to attenuate observed effect sizes (Schmidt & Hunter, 1996), the unreliability of the product term constitutes one more reason to expect that moderator effect sizes will be relatively small and will be detected reliably only with relatively large samples.

In summary, researchers planning studies involving moderator tests should recognize that effect sizes for these tests are likely to be relatively small—probably no larger than $f^2 = .075$ and certainly no larger than $f^2 = .15$. Effect sizes may be still smaller when the interaction is theorized to be disordinal (crossed) and when reliability of X and M is relatively low. Thus, large samples (minimum of $N = 150$ or 200) will probably be necessary to reliably detect moderator effects in most research areas.

Mediator and Moderator Tests in Clinical Trials—MacArthur Group Recommendations

Kraemer, Kiernan, Essex, and Kupfer (2008) discussed modifications recommended by the MacArthur Group (Kraemer et al., 2002) in applying Baron and Kenny's (1986) approaches to clinical trials. In the basic two-group clinical trial, participants are randomized to an intervention group or a comparison group, and the main hypotheses focus on group differences on one or more outcome variables. Because of the focus on group differences, it is natural to analyze data of this type using ANOVA or ANCOVA, but there are advantages to using the more general regression framework for these analyses, particularly when the focus extends beyond treatment differences to search for important intervening variables (mediators) that account for treatment effectiveness or to test (moderator) hypotheses about when and for whom this effectiveness is greatest. When researchers search for mediators and moderators of intervention outcomes, the Baron and Kenny framework provides a valuable starting point, but Kraemer et al. (2008) argued for specific modifications when using mediator and moderator tests in the context of clinical trials.

To include categorical variables (such as type of treatment) in regression analyses, it is necessary to develop a set of numerical code variables that carry all of the category information (i.e., that unambiguously specify the category to which each participant belongs; Cohen et al., 2003, chapter 8). In the two-group case, only one code variable is needed, and it is common to use a dummy coding approach in which the treatment group is coded as 1 and the comparison group is coded as 0. Kraemer et al. (2008) recommended an alternative coding strategy, coding the treatment group as 1/2 and the comparison group as $-1/2$. The main findings (R^2, standardized regression coefficients) will be identical whichever coding strategy is selected (Cohen et al., 2003), but the correct interpretation of many parameter estimates, such as unstandardized regression coefficients and the regression intercept, depends on knowing how categorical variables were coded. Researchers reporting regression results involving categorical variables should be sure that this information is included both in the text and in any tables presenting these findings.

The major modifications proposed by Kraemer et al. (2008) pertain to establishing strong evidence for causation when testing mediator hypotheses and to the interpretation of interaction effects when the (experimentally manipulated) IV is correlated with the putative moderator variable. As noted earlier, it has been common for researchers testing mediation hypotheses to measure *X, M,* and *Y* (see Figure 2) at the same point in time, creating ambiguity about the causal relations among the variables. Stronger evidence of causation is obtained when *X* is measured at an earlier time point than *M* and ideally when *Y* is measured at a later time point than *M* (Maxwell & Cole, 2007). Kraemer et al. noted that the social consequences of misspecifying mediator models in intervention research are potentially large and that the requirement of temporal precedence of *X* over *M* should be routinely imposed in evaluating mediator research. That is, it is important for researchers to show that group differences on the putative mediator variable emerge over time (presumably as a result of intervention) and are not present at the time of group assignment (e.g., as a consequence of randomization failure). This requirement, which goes beyond Baron and Kenny's (1986) recommendations, nonetheless will not be a hardship for most intervention researchers because it is natural to measure *M* at some time during the intervention (or sometimes even at its conclusion).

The second issue raised by Kraemer et al. (2008) specific to the context of intervention research concerns the interpretation of statistical interactions between *X* and *M*. Note that Figure 6 assumes that both *X* and *M* may be related to *Y* (main effects) but that they are not correlated with one another. In research on measured IVs, however, it is not uncommon for *X* and *M* to be correlated. (This would be specified as an additional path, in the form of a two-headed arrow connecting *X* to *M* in Figure 6.) In research on measured IVs, this correlation does not typically alter the interpretation of the association between the product term and the DV as a moderator effect (Baron & Kenny, 1986).

In intervention research, however, a significant correlation between *X* and *M* has different ramifications because *X* is an experimentally manipulated variable, rather than a measured variable. Thus, if *X* and *M* are correlated (i.e., if the groups differ in their mean scores on *M*), we are justified in interpreting this association as a cause–effect relationship: Something about the intervention is causing changes in participants' scores on *M*. Kraemer et al. (2008) argued that such a relationship fits the definition of mediation, rather than moderation, because the variable *M* is a proximal outcome of the intervention that leads to group differences (for at least some values of *M*) on the distal outcome *Y*.

This understanding of the *X–M* association, which stems from the status of *X* as an experimentally manipulated variable, yields three further recommended modifications to the Baron and Kenny (1986) approach when applied in the context of randomized trials. First, when testing moderator relations, *M* should be measured before *X*. This is typically the case in intervention research, as moderator variables would normally be pretreatment scores on variables thought to determine response to intervention. Second, when *M* is presumed to be a moderator of treatment effectiveness, there should be no association between *X* and *M*. This is also typically the case when *M* is measured before the start of intervention because random assignment should result in groups that are comparable with respect to *M*. Third, when testing mediator models, we should not assume that there is no interaction (as is typical in the Baron & Kenny, 1986, method) but rather should include the interaction term in the model (after centering *X* and *M*) because it is possible that the degree of mediation differs as a function of the value of *M*. Because of the causal relation between *X* and *M,* this interaction, if significant, is still best interpreted as mediation rather than moderation, for the reasons just discussed (Kraemer et al., 2008).

Mediated Moderation, Moderated Mediation

Baron and Kenny (1986) also described more complex causal models combining mediation and moderation. In our moderator example, we found a pattern of associations consistent with the hypothesis that social support moderates the effect of pain on catastrophic thinking. When social support is high, Figure 7 suggests that decreasing pain should lead to a decrease in levels of catastrophizing. When social support is low, however, decreasing pain will not be too helpful in decreasing catastrophic thinking. To gain a better understanding of this association, one could look for mediators of this interaction effect. For example, it might be hypothesized that the effect of pain on catastrophizing is mediated by a constricted sense of self. Persons experiencing high levels of pain may tend to identify strongly with the cause of the pain (i.e., with a disability), making it more difficult to envision a positive future and more natural to dwell on potential negative outcomes (i.e., to catastrophize). If lack of close relationships also tends to promote a constricted sense of self (to a large extent independent of level of pain), then this intervening variable may account, at least in part, for the joint relation of pain and social support with catastrophizing that is depicted in Figure 7. Baron and Kenny referred to this pattern of relations as *mediated moderation* because the moderator effect (pain and social support) on the DV is partly explained by a similar joint association with another variable

(i.e., constricted sense of self) that in turn affects the DV. Understanding what variables mediate observed moderator relations can have important implications for both theory and intervention.

When research supports a theorized mediator effect, it may also be useful to consider whether this mediator effect operates similarly for all persons or in all contexts. The question of when or for whom a mediator effect may be stronger or weaker was defined by Baron and Kenny (1986) as a question about *moderated mediation*. There are a number of different ways that a fourth variable can moderate a demonstrated mediator relation (Preacher, Rucker, & Hayes, 2007). Figure 2B shows the three causal paths that characterize a basic mediation. In theory, the strength (and perhaps even the direction) of any of these paths could depend on a person's standing on a fourth variable measuring either a personal characteristic or a contextual factor relevant to the causal processes being studied. Thus, Preacher et al. (2007) referred to moderated mediation hypotheses as questions about conditional indirect effects. Moderated mediation hypotheses take the logic of simple moderator hypotheses (about the conditionality of the association between X and Y) and extend it to a more complex network of hypothesized causal linkages (between X and Y and a hypothesized mediator M).

Preacher et al. (2007) provided detailed conceptual and statistical guidelines for testing moderated mediation hypotheses, as well as links to computer methods to make these techniques more accessible to researchers. Muller, Judd, and Yzerbyt (2005) showed that mediated moderation and moderated mediation, although conceptually distinct, have considerable overlap in terms of their underlying structural model and offered helpful guidelines for analyzing and interpreting both types of hypotheses.

Effect Sizes and Confidence Intervals in MRC

Researchers in psychology are advised to "always present effect sizes for primary outcomes" and that "interval estimates should be given for any effect sizes involving primary outcomes" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599). In addition, graphical representations of findings should include interval estimates "whenever possible" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 601). The first of these injunctions is easy to follow for researchers who conduct their primary analyses using MRC, when effect size is understood in the general sense of a numeric estimate reflecting the "strength of relationship" (APA, 2001, p. 26) between two or more variables. Regression coefficients (B or β) reflect the magnitude of a presumed causal relation between two variables, the squared multiple correlation coefficient (R^2) reflects the proportion of variance in the DV that is explained by a set of predictor variables, and the squared semipartial correlation coefficient (sr^2) estimates the proportion of DV variance that is uniquely explained by a given predictor variable, controlling for the other IVs in the equation. The choice of which effect size is most informative will depend on the nature of the research question (e.g., tests of causal hypotheses vs. purely predictive analyses) and on the scaling of the IVs and the DV (i.e., on whether the units of measurement are intuitively meaningful or not; Hoyt, Leierer, & Millington, 2006).

The recommendation to report interval estimates, or CIs, along with effect sizes has yet to be widely adopted in psychology journals. Although acknowledging that effect sizes and CIs "are, in general, the best reporting strategy" (APA, 2001, p. 22), the fifth edition of the *Publication Manual* provides no examples of how CIs should be reported in article text or tables, much less how they can be usefully included in figures. Unlike test statistics and p values, CIs are not part of the routine output of most common statistical software packages, so they require additional calculation on the part of researchers. For some effect sizes commonly reported in MRC analyses, the CI is usually asymmetrical and somewhat complicated to compute. In the following sections, we briefly discuss how CIs can assist authors and readers with accurate interpretations of research findings, present general principles for computing CIs, show how these apply to the four most common effect sizes reported in MRC analyses, and mention several less common cases that may also be useful to researchers making use of MRC.

Utility of CIs for Psychological Research

Reporting of effect sizes and CIs has been recommended as an alternative to the near-universal reliance on null hypothesis significance testing (NHST) that characterized psychological research in the latter half of the 20th century (Kline, 2004). Critics argue that reports of NHSTs are frequently misunderstood even by researchers with statistical training (Hunter, 1997) and that they tend to obscure study findings by encouraging attention to p values or even number of asterisks reported in tables, rather than the actual effect size—to statistical significance rather than the practical significance of the relations under investigation (Kirk, 1996). Researchers or readers focusing on p values are apt to forget that p is partly a function of N, so that significant p values will be obtained for small or even trivial effect sizes, given a large enough sample (Cohen, 1994). Reliance on p values (or dichotomous results of statistical significance tests) rather than effect sizes can also undermine efforts to cumulate findings across studies—a key goal of scientific inquiry (Schmidt, 1992).

Once the utility of reporting effect-size estimates has been accepted, augmenting this information with CIs, which specify the precision of these estimates, is a natural next step. CIs remind readers that sample parameter estimates will differ from population values because of sampling error and give a sense of the variability that would be expected in these estimates if the study in question were replicated multiple times. CIs can provide the same evidence as NHSTs and can also assist readers to combine evidence across studies. They encourage meta-analytic thinking focused on synthesis of evidence from multiple sources about the strength of association between variables in the population (Cumming & Finch, 2005). Finally, evidence on precision of effect-size estimates may be more useful for purposes of research design and interpretation than calculations of statistical power (Cohen et al., 2003; Kelley & Maxwell, 2003).

How Are CIs Computed?

Good primers on CIs for a variety of different types of effect sizes include Cumming and Finch (2005) and Kline (2004). In this brief summary, we follow the notation used by Cohen et al. (2003, pp. 86–88). Suppose that we are interested in the direct effect of pain intensity (X_1) on catastrophizing (Y), controlling for stress (X_2) and activity interference (X_3), as in Figure 3. When Y is regressed simultaneously onto the three predictor variables, the unstandardized partial regression coefficient for X_1 (i.e., $B_{Y1\cdot 23} = 0.064$) and its standard error ($SE_B = 0.024$) are part of the standard output for the analysis. To compute the appropriate CI, we must specify the desired level of confidence (C). For this example (and throughout this article), we set $C = .95$ (a 95% CI), reflecting the conventional alpha level of .05. However, arguments can be made for reporting on lower levels of precision, such as $C = .80$ or even $C = .67$ (Cohen et al., 2003, pp. 86–88).

Given C and the sample size ($N = 143$ for this example), we can compute the critical value on the t distribution corresponding to our chosen confidence level ($t_C = 1.977$). The margin of error is equal to the product of SE_B and t_C:

$$me = (t_C)(SE_B). \qquad (2)$$

For our example, $me = (1.977)(0.024) = 0.047$. The margin of error represents half the width of the CI (i.e., the CI extends for *me* units on either side of B). The lower and upper limits of the CI are computed by subtracting *me* from B and adding *me* to B, respectively:

$$\mathrm{CL}_{\mathrm{lower}} = B - me, \text{ and} \qquad (3)$$

$$\mathrm{CL}_{\mathrm{upper}} = B + me, \qquad (4)$$

where $\mathrm{CL}_{\mathrm{lower}}$ and $\mathrm{CL}_{\mathrm{upper}}$ are the lower and upper limits of the CI. Thus, for our example, $\mathrm{CL}_{\mathrm{lower}} = 0.064 - 0.047 = 0.017$, and $\mathrm{CL}_{\mathrm{upper}} = 0.064 + 0.047 = 0.111$. To report this finding in text, we would state that the regression coefficient for social support in this equation was $B = 0.064$ (95% CI = 0.017, 0.111). (Note that it is conventional to omit the subscripts from B in the research report, which means that it is important to include a verbal description of what other predictor variables were included in the analysis whenever effect sizes are reported in MRC.)

In this example, B is referred to as the *point estimate* or *parameter estimate,* and the CI gives an indication of the precision of this estimate. The CI defines a range of plausible values for the population parameter (i.e., the unstandardized regression coefficient in the population). Cumming and Finch (2005) suggested a reasonable interpretation of this interval: “We can be 95% confident that our CI includes [the population value of B]” (p. 175). Thus, although our sample B is 0.064, this finding is consistent with an actual (population) value of B as small as 0.02 and as large as 0.11, which reminds us to be cautious in interpreting the observed effect size. In addition, the fact that the 95% CI does not include zero indicates that the observed value of B differs significantly from zero ($p < .05$).

Because this method for reporting CIs involves some calculation based on standard output, we offer two recommendations that may be helpful. (We make additional suggestions about available software later on.) First, it is a good idea to automate this process as much as possible. When performing hand calculations, spreadsheet software such as Microsoft Excel is preferable by far to using a hand calculator. Relevant quantities (i.e., B, SE_B, and df) can be entered into cells in the spreadsheet, with formulas based on these cell values for computing the critical value of t, the *me,* and the upper and lower CLs. Readers comfortable with the use of syntax can create simple functions (in R or S-Plus) or macros (in SPSS or SAS) to compute CIs from these same input values.

A second consideration is the number of significant digits necessary to compute an accurate CI. When the focal predictor variable has a much greater range than the DV (i.e., when $SD_X > SD_Y$), the unstandardized regression coefficient is likely to be small (because a one-unit change in X represents a relatively small change relative to the range of X and must predict a very small change in the raw Y score). In many statistical packages, the default is to report three decimal places, which may result in some rounding error if these default values are used to compute a CI. In such cases, it is advisable to obtain more precise estimates of B and SE_B for computing the CI. For example, the unstandardized regression coefficient for the product term in Table 2 is $B = .005$, and its standard error was reported in the SPSS output as $SE_B = .002$. By double-clicking on these values in the SPSS pivot tables, one can obtain the more precise values of these parameter estimates, which we rounded to three significant digits (i.e., $B = .00533$ and $SE_B = .00198$) for computation of the corresponding CI.

CIs for Standardized Effect Sizes

Unfortunately, this straightforward procedure for computing CIs yields incorrect interval estimates for many of the effect sizes commonly reported in MRC analyses. In particular, standardized coefficients such as Pearson’s r, multiple R^2, the standardized regression coefficient (β), and the squared semipartial coefficient (sr^2) all have asymmetric CIs in most circumstances, so that CIs should be estimated using special computational procedures.

Pearson’s r. The asymmetry of the sampling distribution for Pearson’s r is well known, and Fisher’s r to z' transformation is the recommended method for deriving CIs for correlation coefficients or for averaging correlations across studies (Hedges & Olkin, 1985). As described by Cohen et al. (2003, pp. 45–46), the procedure involves converting r to z' using the formula

$$z' = (.5)[\ln(1 + r) - \ln(1 - r)]. \quad (5)$$

The standard error of z' depends only on the sample size and is equal to $1/\sqrt{\mathrm{N} - 3}$. CIs for z' are symmetric and can be computed using the standard method described above, but using the critical value of the normal (z) distribution rather than the t distribution. When the upper and lower confidence limits for z' are converted back to the corresponding values of r (by reversing the procedure of Equation 8), the resulting values give an accurate (and asymmetric) CI around the original r.

Multiple R^2. Steiger and Fouladi (1997) described a procedure for deriving a CI for R^2 using the noncentral F distribution. The upper and lower confidence limits are not directly computable but can be determined by a process of trial-and-error estimation, which can be automated using any statistical software package that incorporates the family of noncentral F distributions. Smithson (2001) created freely available SPSS scripts to compute accurate CIs for R^2. These scripts may also be used to compute CIs for pR^2 (i.e., the squared multiple partial correlation coefficient) but unfortunately not for sR^2 (the squared multiple semipartial correlation, which is here called ΔR^2).

Other standardized effect sizes. An important recent innovation is the development of MBESS (Methods for the Behavioral, Educational, and Social Sciences; Kelley, 2007a, 2007b), a package of functions for the open-source R statistical program. MBESS includes a number of functions to compute CIs for standardized effect sizes using the noncentral t, F, and χ^2 distributions. Researchers not familiar with R can easily use statistical output from more familiar software programs as input for these functions, making accurate CIs for regression effect sizes available without the need for extensive hand calculation or knowledge of statistical programming languages. MBESS is relatively new and is still under development, and new functions will likely continue to be added. The accessibility of MBESS, which can be freely downloaded, along with the R base system, from the R Project Web site (www.r-project.org), represents a large step toward making the ideal of routine reporting of effect sizes and CIs in psychological research a reality.

CIs for Comparing Correlations or Regression Coefficients

A common error in interpreting regression and correlational analyses is to overinterpret observed differences between observed effect sizes. For example, when one of two predictors in a regression equation has a higher standardized regression coefficient, it is only natural to conclude that that variable is more important in explaining variance on the DV. Similarly, when examining a correlation matrix, it is natural to conclude that the small differences in the magnitude of correlation coefficients reflect important differences in strength of association among the variables.

One benefit of reporting CIs is that it reminds us just how imprecise effect-size estimates (typically) are in psychological research. For example, in Table 1, the largest correlation (between catastrophizing and activity interference) is $r = .52$, but the 95% CI spans a range that includes nearly every other correlation in the table. This serves as a reminder that, if we were to replicate this study, the pattern of correlations in our second data set might look quite different from the pattern in Table 1.

A counterintuitive corollary is that just because one effect size is statistically significant and the other is not, it is not legitimate to conclude that the difference between the two is statistically significant. For example, notice that the CI for the only nonsignificant correlation in Table 1 (r_{14}, between pain and social support) overlaps with that for r_{45} (between social support and activity level). Despite the fact that one is significantly different from zero and the other is not, the overlap in their CIs suggests we should be cautious about concluding on this basis that activity interference is more strongly related to social support than is pain.

Comparative conclusions such as these need to be grounded in statistical tests for the differences between correlations (e.g., Cohen et al., 2003, p. 49) or regression coefficients (e.g., Azen & Budescu, 2003). Alternatively, researchers can construct CIs for the difference between two effect sizes. Zou (2007) has recently developed a general method for comparing correlation coefficients (including multiple R^2) taking into account the asymmetry of the confidence limits around individual effect sizes and has shown that this method has statistical properties superior to available alternatives. Computing CIs for these differences is desirable for the same reason that CIs are useful in general—they provide information about the likelihood that the true difference could be zero but also give easily interpretable information about the precision of this difference estimate.

Summary: CIs and MRC

Researchers who publish findings using MRC have been slow to adopt the recommendation of the methodological community to report effect sizes and CIs for primary findings. Effect sizes are readily available in MRC, but the confidence limits for most of these are not at all straightforward to compute, and the most commonly used statistical software packages have done little to ease this burden. In the past 10 years, however, significant progress has been made in providing the tools needed to report CIs for most MRC effect sizes. CIs for r and B are most readily obtained, although, in each case, researchers must make additional calculations from the output generated by standard statistical packages. CIs for other standardized effect sizes and for comparisons between effect sizes are often obtainable but require specialized applications that may be daunting to many researchers using MRC. However, the progress in the past decade is very encouraging, and increasingly accessible tools will in all probability continue to be developed. We encourage researchers making use of MRC to familiarize themselves with the meaning and uses of CIs and to begin moving the field toward this next generation of statistical reporting practices.

Other Important Topics

In this article, we have focused on the linkage between analysis and theory and on improving reporting and interpretation of findings from theory-derived regression analyses. Topics not emphasized in our presentation but critical to making best use of MRC techniques include statistical assumptions underlying significance testing (and CIs) in MRC, considerations of statistical power and precision in determining sample sizes, techniques for addressing missing data, and the effect of measurement error on interpretation of findings. We briefly introduce each of these issues here, referring interested readers to thorough treatments in Cohen et al. (2003) and to an excellent overview (with references to more detailed resources on specific topics) by Kelley and Maxwell (in press), to be published in the forthcoming *Quantitative Methods in the Social and Behavioral Sciences: A Guide for Researchers and Reviewers* (Hancock & Mueller, in press).

Statistical Assumptions in MRC

Significance tests and standard errors (and therefore CIs) in MRC are predicated on assumptions of normality and heteroscedasticity of residuals (i.e., errors of prediction $Y_i - \hat{Y}_i$). Validity of these assumptions is evaluated by visual examination of a histogram of the residual scores, and of a graph plotting values of the predicted scores (on the *x*-axis) against the corresponding residuals (*y*-axis).

Other critical assumptions that can bias effect-size estimates as well as significance tests include independence of observations and correct model specification. Independence of observations may be violated if participants fall into naturally occurring groups (e.g., families, treatment settings) where a person's outcomes are likely to be more similar to those of another in the same group than they are to those of a person in a different group. Grouping variables may also be created in the course of the study, as when participants receive group interventions or when clients receiving individual interventions are nested within therapists. Kenny et al. (1998) described the biasing effects of nonindependence on effect-size estimates in different research designs, as well as procedures for appropriate analysis of nested data.

Correct model specification is perhaps the most important and most underappreciated assumption underlying interpretation of findings in MRC. This is especially true when MRC is applied to observational data, so that there is no logical or empirical warrant for inferences about causal relations and no control for potential confounding variables that may be important in understanding the associations under investigation. This assumption reinforces the importance, highlighted throughout this article, of careful theoretical justification for proposed causal hypotheses, for exploration of plausible alternative structural models for the observed variables, and, above all, for thorough consideration of unmeasured (i.e., omitted) variables that may be important in understanding the observed relationships.

Power and Precision

Statistical power should always be a factor in designing research investigations because inquiries in many areas of the social sciences have been chronically underpowered (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Hoyt, Leierer, and Millington (2006) provided a brief introduction to power analysis for MRC. In keeping with a focus on magnitude-of-effect estimates and their precision (CIs), several methodologists have lately turned attention to precision of effect-size estimates as a consideration for sample-size calculations (Cohen et al., 2003; Kelley & Maxwell, 2003). In designing studies, investigators should decide whether power (of statistical significance tests) or precision (of effect-size estimates) is the more important factor for determining sample size.

Missing Data

Observational studies, especially those that are longitudinal in design, usually are subject to some amount of missing data, varying from occasional omissions of a question from a larger questionnaire to missing scores on an entire questionnaire to missing scores on all questionnaires on one or more measurement occasions. Decisions about how to address missing data are beyond the scope of this article but are addressed in some detail by Cohen et al. (2003), with a helpful overview by Kelley and Maxwell (in press). The researcher's focus should be to address missing data in a way that does not bias findings and that maximizes power and precision.

Measurement Error

Hoyt, Leierer, and Millington (2006) reviewed factors affecting magnitude-of-effect estimates in MRC. A pervasive reality in observational research is measurement error, which attenuates standardized coefficients reflecting bivariate associations and has more complex effects on standardized partial coefficients (reflecting associations between *X* and *Y* controlling for other predictor variables). Hoyt, Warbasse, and Chu (2006) noted that measurement error may be systematic as well as random and addressed the issue of *construct-irrelevant variance* (i.e., variance in scores that reflects participant characteristics other than the construct of interest) as a source of bias in psychological research. The implications of this analysis are that researchers need to pay attention to issues of construct validity as well as reliability in selecting research measures. Validity evidence should be reported routinely and considered with respect to its implications for interpretation of findings.

Conclusions

MRC techniques are valuable both in their own right and as a bridge to more sophisticated model-testing approaches such as structural equation modeling and multilevel modeling. To exploit the full potential of MRC, researchers should start with a sound knowledge of theory in the research area and develop research hypotheses that will support or refute theory. Careful attention to the match between hypothesis and research design (including selection of measures) and analysis will avert many common problems found in MRC research reports, including overinclusion of covariates, focus on prediction rather than explanation in interpreting regression findings, and overuse of empirical (stepwise) approaches. Mediator and moderator tests, although still subject to some degree of controversy and ongoing refinement, are well-established techniques for elaborating theory-derived associations among networks of variables. However, researchers fail to reap the potential of these approaches when they limit themselves to cross-sectional data and when they are less than assiduous in their consideration of plausible rival explanations for their findings. Finally, it is important to move MRC research, as well as research using other analytic tools, away from exclusive reliance on significance testing as the lens through which findings are interpreted. MRC fosters attention to effect sizes, and researchers should bracket these with confidence limits when possible to remind themselves and their readers of the importance of precision of effect-size estimation as well as statistical power.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Aiken, L. S., & West, S. G. (2000). Multiple regression. In *Encyclopedia of psychology* (Vol. 5, pp. 350–352). New York: Oxford University Press.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8,* 129–148.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Banks, S., & Kerns, R. (1996). Explaining high rates of depression in chronic pain: A diathesis–stress framework. *Psychological Bulletin, 119,* 95–110.

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.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40,* 373–400.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects.* New York: Harper & Row.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

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.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

Duncan, O. D. (1975). *Introduction to structural equation models.* New York: Academic Press.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51,* 115–134.

Geisser, M. E., Robinson, M. E., & Riley, J. L. (1999). Pain beliefs, coping and adjustment to chronic pain: Let's focus more on the negative. *Pain, 8,* 161–168.

Hancock, G. R., & Mueller, R. O. (Eds.). (in press). *Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers.* Mahwah, NJ: Erlbaum.

Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hoyt, W. T., Leierer, S., & Millington, M. J. (2006). Analysis and interpretation of findings using multiple regression techniques. *Rehabilitation Counseling Bulletin, 49,* 223–233.

Hoyt, W. T., Warbasse, L. E., & Chu, E. Y. (2006). Construct validation in counseling psychology research. *The Counseling Psychologist, 34,* 769–805.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7.

International Association for the Study of Pain. (1986). Classification of chronic pain: Descriptions of chronic pain syndromes and definitions of pain terms—The International Association for the Study of Pain, Subcommittee on Taxonomy. *Pain, 3*(Suppl.), S1–S226.

Jaccard, J., Guilamo-Ramos, V., Johansson, M., & Bouris, A. (2006). Multiple regression analyses in clinical child and adolescent psychology. *Journal of Clinical Child and Adolescent Psychology, 35,* 456–479.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs, 1,* 57–93.

Jones, D. A., Rollman, G. B., White, K. P., Hill, M. L., & Brooke, R. I. (2003). The relationship between cognitive appraisal, affect, and catastrophizing in patients with chronic pain. *The Journal of Pain, 4,* 267–277.

Judd, C., & McClelland, G. (1989). *Data analysis: A model comparison approach.* New York: Harcourt Brace Jovanovich.

Keefe, F. J., Brown, G. K., Wallston, K. A., & Caldewell, D. S. (1989). Coping with rheumatoid arthritis: Catastrophizing as a maladaptive strategy. *Pain, 37,* 51–56.

Keefe, F. J., Rumble, M. E., Scipio, C. D., Giordano, L. A., & Perri, L. M. (2004). Psychological aspects of persistent pain: Current state of the science. *The Journal of Pain, 5,* 195–211.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20,* 1–24.

Kelley, K. (2007b). Methods for the Behavioral, Educational, and Social Sciences: An R package. *Behavior Research Methods, 39,* 979–984.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321.

Kelley, K., & Maxwell, S. E. (in press). Multiple regression: A guide for researchers and reviewers. In G. R. Hancock & R. O. Mueller (Eds.), *Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers.* Mahwah, NJ: Erlbaum.

Kenny, D. A. (1979). *Correlation and causality.* New York: Wiley.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In *Handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). New York: McGraw-Hill.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology, 27*(Suppl.), S101–S108.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59,* 877–884.

Kwok, O.-M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology, 53,* 370–386.

Lee, G. K., Chan, F., & Berven, N. L. (2007). Factors affecting depression among people with chronic musculoskeletal pain: A structural equation model. *Rehabilitation Psychology, 52,* 33–43.

Lewinsohn, P. M., Hoberman, H. M., Teri, L., & Hautzinger, M. (1985). An integrative theory of depression. In S. Reiss & R. R. Bootzin (Eds.), *Theoretical issues in behavior therapy* (pp. 331–359). New York: Academic Press.

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., 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.

Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology, 53,* 372–378.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12,* 23–44.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113,* 181–190.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McCullough, M. E., Hoyt, W. T., Larson, D. B., Koenig, H., & Thoresen, C. (2000). Religious involvement and mortality: A meta-analytic review. *Health Psychology, 19,* 211–222.

Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology, 77,* 143–148.

Mueller, R. O. (1997). Structural equation modeling: Back to basics. *Structural Equation Modeling, 4,* 353–369.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89,* 852–863.

Nielsen, M. S. (2003). Prevalence of posttraumatic stress disorder in persons with spinal cord injuries: The mediating effect of social support. *Rehabilitation Psychology, 48,* 289–295.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). Fort Worth, TX: Harcourt Brace.

Petrocelli, J. V. (2003). Hierarchical multiple regression in counseling research: Common problems and possible remedies. *Measurement and Evaluation in Counseling and Development, 36,* 9–22.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36,* 717–731.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42,* 185–227.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422–445.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605–632.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290–312). San Francisco: Jossey-Bass.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical methods. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

Turk, D. C. (1999). Continuity and change. *Clinical Journal of Pain, 15,* 163–165.

Turk, D. C. (2003). Chronic pain and whiplash associated disorders: Rehabilitation and secondary prevention. *Pain Research and Management, 8,* 40–43.

Turk, D. C., & Okifuji, A. (2002). Psychological factors in chronic pain: Evolution and revolution. *Journal of Consulting and Clinical Psychology, 70,* 678–690.

Turner, J. A., Mancl, L., & Aaron, L. A. (2004). Pain-related catastrophizing: A daily process study. *Pain, 110,* 103–111.

Wampold, B. E., Davis, B. & Good, R. H. (1990). Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology, 58,* 360–367.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425–433.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wolfle, L. M. (1985). Applications of causal models in higher education. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 1, pp. 381–413). New York: Agathon.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12,* 399–413.

Received May 19, 2008
Revision received June 11, 2008
Accepted June 11, 2008 ■