Journal of Family Psychology
2005, Vol. 19, No. 1, 111–120

0893-3200/05/$12.00 DOI: 10.1037/0893-3200.19.1.111

Designing, Testing, and Interpreting Interactions and Moderator Effects in Family Research

Mark A. Whisman and Gary H. McClelland
University of Colorado at Boulder

This article is a primer on issues in designing, testing, and interpreting interaction or moderator effects in research on family psychology. The first section focuses on procedures for testing and interpreting simple effects and interactions, as well as common errors in testing moderators (e.g., testing differences among subgroup correlations, omitting components of products, and using median splits). The second section, devoted to difficulties in detecting interactions, covers such topics as statistical power, measurement error, distribution of variables, and mathematical constraints of ordinal interactions. The third section, devoted to design issues, focuses on recommendations such as including reliable measures, enhancing statistical power, and oversampling extreme scores. The topics covered should aid understanding of existing moderator research as well as improve future research on interaction effects.

Keywords: interaction, moderator, moderation

There are many examples of theories in family psychology in which the association between two variables is hypothesized to be dependent on some other variable. Studies that evaluate these theories have evaluated interactions and moderator effects.[1] For example, a perusal of articles published in the *Journal of Family Psychology* in 2003 (Volume 17) indicates that regression analyses have been used to study interaction effects in basic research evaluating outcome variables such as marital violence, children's health and adjustment, attitudes about physical punishment, adolescent mothers' psychological adjustment, and quality of parenting. In addition to basic research on family functioning, moderator effects have also been used to study predictors of differential responses to therapy interventions. The importance of moderator or interaction research was underscored by Cohen, Cohen, West, and Aiken (2003), according to whom "it is safe to say that the testing of interactions is at the very heart of theory testing in the social sciences" (p. 255). However, others (e.g., Judd & McClelland, 1989, 1998; Luce, 1995) warn that interactions are often signs of trouble, indicating scaling problems, model misspecifications, and other difficulties. In any case, it is important to be able to detect and interpret interactions.

Despite the popularity of studying moderator effects in family research, there are widespread misunderstandings regarding the appropriate methods for testing and interpreting interactions. This article presents a primer on issues involved in designing, testing, and interpreting studies evaluating interaction effects. We focus on providing a practical, "hands-on" discussion of interaction effects, using examples from the family literature. In providing examples, we have chosen to focus on appropriate applications of tests for interaction effects from this journal, rather than inappropriate or improper tests for interactions. The article is divided into three major sections: testing and interpreting interactions, difficulties in detecting interactions, and recommendations regarding design issues in conducting studies evaluating interaction effects.

Testing and Interpreting Interactions

How to Test for Interactions

As a context for describing appropriate statistical methods for testing for interaction effects, we consider as a prototypical example testing for the ability of family support (F) to moderate the relationship between life event stress (L) and depression (D).[2] A reasonable model is that life event stress is likely to lead to depression but that strong family support might buffer individuals against the effects of stress. The test of the moderating or buffering effect of family support is the statistical comparison of the following two models:

Several applets illustrating some of the issues discussed in this article are available at http://psych.colorado.edu/~mcclella/JFP/applets/.

Correspondence concerning this article should be addressed to Mark A. Whisman or Gary H. McClelland, Department of Psychology, University of Colorado at Boulder, 345 UCB, Boulder, CO 80309-0345. E-mail: whisman@colorado.edu or gary.mcclelland@colorado.edu

[1] We use the terms *interaction* and *moderation* synonymously.

[2] We assume that the criterion variable is continuous. Issues involved in testing and interpreting interactions in logistic regression when Y is dichotomous are similar to those discussed here, but there are special issues pertaining to logistic regression; see Jaccard (2001) for practical advice.

$$\text{Additive: } D = b_0 + b_1L + b_2F$$

and

$$\text{Moderator: } D = b_0 + b_1L + b_2F + b_3L \times F.$$

The last term in the moderator model is simply the product of the two predictors. There are a number of equivalent methods for testing whether the difference between the two models is statistically significant. Three of the most common are (a) testing whether the increment in the squared multiple correlation (ΔR^2) is significantly greater than zero, (b) testing whether the coefficient b_3 differs from 0, and (c) testing whether the partial correlation between the product $L \times F$ and *D,* when controlling for *L* and *F,* differs from 0. The same test for the moderating effect of family support is appropriate regardless of whether it is measured on a continuous scale or whether it is a dichotomous predictor (e.g., whether a person has a spouse or not). In the latter case, one must code *F* using, for example, 1, 0 dummy codes or (often more usefully) $\frac{1}{2}, -\frac{1}{2}$ contrast codes. We consider the relative advantages of different coding schemes later, but the test of the moderating effect is independent of what coding is used. The method just described is simple—it can be implemented with any multiple regression program—and is the only appropriate statistical method for testing the significance of a moderator effect (Aiken & West, 1991; Cohen, 1968, 1978; Jaccard & Turrisi, 2003; Judd & McClelland, 1989). Before turning to a discussion of how to interpret the coefficients of the moderator model, we first consider some common errors encountered in testing this model.

Testing differences among subgroup correlations. It is tempting to believe that if *F* moderates the relationship between *L* and *D,* then there must be differences among the correlations between *L* and *D* computed at different levels of *F* or for different subgroups defined by *F.* If so, then moderation could be assessed by testing the differences between these subgroup correlations. However, there is no necessary relationship between the presence or absence of a significant moderator and whether or not subgroup correlations would be significantly different. The problem is that correlations confound both true moderating effects with differences in predictor variance. Figure 1a shows an example in which the relationship between *L* and *D* (depicted by solid and dashed lines) is identical (both slopes = 0.58) for two subgroups defined by *F* (depicted by open and solid circles); hence, there is no moderating effect of *F.* However, because the variability of *L* is much smaller in one group than the other, perhaps owing to a restriction in its range for one of the subgroups, the correlations of .38 and .70 are significantly different. Figure 1b shows the opposite example in which there is a moderating effect of *F* (two slopes of 0.98 and 0.41), but, because of the greater variability of the predictor *L* in the group with the lower slope, the correlations are both .57. This is not to say that one should never test for differences in subgroup correlations. If this is the relevant question, it is the question that should be addressed. However, such tests should not be called "moderation tests"

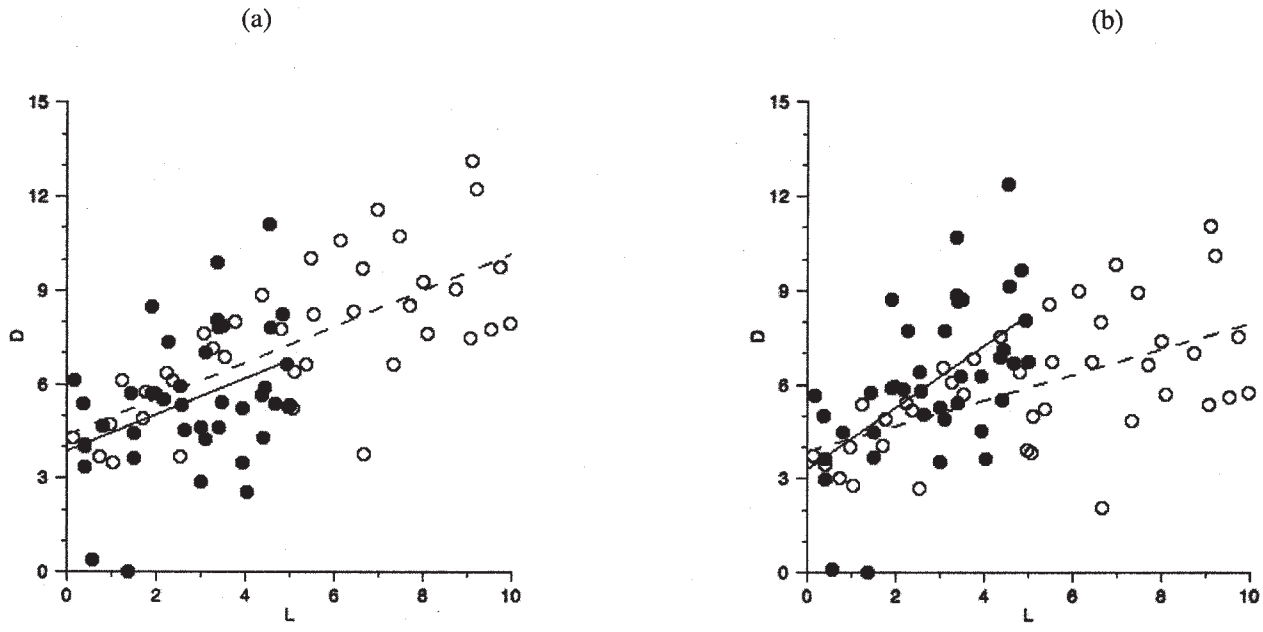

Figure 1. a: Hypothetical data illustrating that a difference in correlation does not imply moderation. Correlations for the closed and open data points are .38 and .70, respectively; solid and dashed lines are the respective best-fitting regression lines, both with a slope of .58. b: Hypothetical data illustrating that moderation (difference in slopes, .98 and .41) does not imply a difference in correlations (both *r*s = .57).

and should never be confused with the appropriate test of the interaction effect as described earlier.

Omitting components of products. A common error when testing the moderation regression model is to include the product ($L \times F$ in our example) while failing to include both of the individual components (L and F in our example). Cohen (1978) emphasized that the product represents the interaction only when its components have been partialed out. Leaving out the individual components in the regression model inherently confounds the additive and multiplicative effects, producing biased and misleading results. The simple rule is that the components of any products must always be included when testing the moderator effect.

Median splits. The most egregious of the common errors made in testing moderators is to form groups by splitting each predictor variable at its median and then performing a two-way analysis of variance (ANOVA). At one time, testing of interactions was always taught in ANOVA courses but seldom in courses on multiple regression. Median splits became a popular method for converting a regression analysis into an ANOVA so that interactions could be tested. In their critique, West, Aiken, and Krull (1996) referred to this approach as "ANOVA with cutpoints." Now that a number of sources (e.g., Aiken & West, 1991; Jaccard & Turrisi, 2003; Judd & McClelland, 1989) provide textbook instruction on how to test and interpret interactions in multiple regression analyses, it is no longer necessary to resort to median splits. More important, median splits are to be avoided because they have a number of deleterious effects. First, median splits are necessarily dependent on particular samples. Researchers studying the same phenomenon may obtain different results simply because the medians happened to be different in their respective samples, and so their variables were split at different points.

Second, in all analyses of the statistical effects of median splits (e.g., Cohen, 1983; Cohen & Cohen, 1983; Irwin & McClelland, 2003; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993), the conclusion has been that they drastically reduce statistical power and squared correlations. Statistical power is already reduced, for reasons discussed subsequently, when searching for moderators in nonlaboratory data (McClelland & Judd, 1993), so any further reductions are highly undesirable. Cohen and Cohen (1983, p. 309), probably the most widely cited authorities on the use of multiple regression in the social sciences, forcefully stated:

> When one reduces a graduated many-valued scale to a two-point scale, one is willfully throwing away information. This has immediate negative consequences to the amount of variance such a crippled variable can account for and concomitantly to the power of the statistical test of its contribution.

A general rule of thumb is that dichotomizing normally distributed predictors reduces the squared multiple correlation to about 64% of what it otherwise would have been (Cohen, 1983), and the reduction is about the same even for highly skewed and bimodal distributions (Irwin & McClelland, 2003). Few family psychology researchers have such strong effects that they can afford these kinds of reductions in their squared correlations. Third, and most important, splitting two or more variables in a multiple regression disrupts the latent covariance structure, resulting in spurious effects, false tests of mediation, and other misleading results (Maxwell & Delaney, 1993; McClelland, Irwin, & Judd, 2004; Vargha, Rudas, Delaney, & Maxwell, 1996). In simple regression, median splits "only" reduce statistical power; in multiple regression, however, especially moderated multiple regression, median splits yield incorrect results. In short, median splits have no place in scientific analyses of moderating effects.

How to Interpret Interaction Effects

Misunderstandings about the meaning of regression coefficients in the moderator model and difficulties in explaining moderating effects have also caused some researchers to resort to median splits. However, this is not necessary, because there is an easy strategy for interpreting regression coefficients in moderator models. Consider the moderation model: $D = b_0 + b_1L + b_2F + b_3L \times F$. If we simply rearrange and regroup the terms, we obtain $D = (b_0 + b_2F) + (b_1 + b_3F)L$, which describes the "simple" relationship between L and D. That is, this equation describes the line relating the two variables for any fixed value of F. The term in the first set of parentheses represents the intercept, and the term in the second set of parentheses represents the regression slope. The interesting part is that both the intercept and the slope depend on the level of F.[3] As F changes, so too do the intercept and the slope of the relationship between L and D. As an example, suppose the estimated moderated regression model was $\hat{D} = 0.5 + 0.8L - 0.05F - 0.16L \times F$, which we can rearrange to $\hat{D} = (0.5 - 0.05F) + (0.8 - 0.16F)L$. This defines an equation relating L and D at each level of F. Suppose that F is measured on a 0 to 5 scale and that L ranges from 0 to 7. Then we can compute the predicted line for each level of F:

$$\hat{D} = 0.50 + 0.80L \text{ if } F = 0$$

$$\hat{D} = 0.45 + 0.64L \text{ if } F = 1$$

$$\hat{D} = 0.40 + 0.48L \text{ if } F = 2$$

$$\hat{D} = 0.35 + 0.32L \text{ if } F = 3$$

$$\hat{D} = 0.30 + 0.16L \text{ if } F = 4$$

$$\hat{D} = 0.25 + 0.00L \text{ if } F = 5$$

[3] The quadratic model $Y = b_0 + b_1X + b_2X^2$ is a special case of the interaction model in which the effect of the predictor variable depends on its own level; in other words, the level of X moderates the relationship between X and Y. Hence, much of our discussion of testing and interpreting interactions, as well as the difficulties of detecting interactions, applies to quadratic models as well. However, the decomposition into simple slopes is more complex and requires finding the derivative; see Judd and McClelland (1989, chap. 10) for details, as well as Aiken and West (1991, chap. 5).

These hypothetical data are consistent with the model that family support buffers the individual against the negative effects of life stress events. According to the first equation, when family support is low or nonexistent (i.e., $F = 0$), the slope (0.80) relating L to D is steep. However, when family support is strongest (i.e., $F = 5$), the slope relating L to D is zero, indicating complete buffering. Plotting the predicted regression lines for each level of F, as in Figure 2, communicates the nature of the moderating effect. In this case, as family support increases, the strength of the relationship between L and D decreases. Although a graph such as that of Figure 2 is usually sufficient, we can be more precise about the exact interpretation of the moderator regression coefficient $b_3 = -0.16$. For each unit increase in F, the slope relating L to D decreases by 0.16.

The strategy for interpreting moderator regression models is thus easy: Plot separate regression lines relating one of the independent variables to the dependent variable for each level of the other independent variable. Either independent variable may be chosen to be the x-axis; usually, the substantive focus of the research will make one choice more useful than the other. If there are too many levels of F to be plotted reasonably, then special values of F such as the mean and one standard deviation above and below the mean provide useful displays of the moderator regression model (for a recent appropriate example of plotting regression lines for specific values, see Liew et al., 2003). Note that calculating a median split based on F and then computing separate regressions within each group is *not* a useful strategy.

Interpreting b_1 and b_2. Nothing seems to cause as much confusion in the testing and interpreting of moderator regression models as the proper interpretation of the coefficients of the individual terms of variables included in products (Irwin & McClelland, 2001). The appropriate interpretation of b_1 is easy to see in the reexpressed form of the moderator model: $D = (b_0 + b_2F) + (b_1 + b_3F)L$. The term $(b_1 + b_3F)$ describes how the slope of the relationship between L and D depends on F. That slope will equal b_1 if and only if $F = 0$. Hence, a test of whether $b_1 = 0$ is a test of the "simple effect" of L on D when and only when $F = 0$; this is not a test of the average effect, so it is most emphatically *not* the "main effect" or the "linear effect" of L on D, as erroneously claimed by far too many researchers.

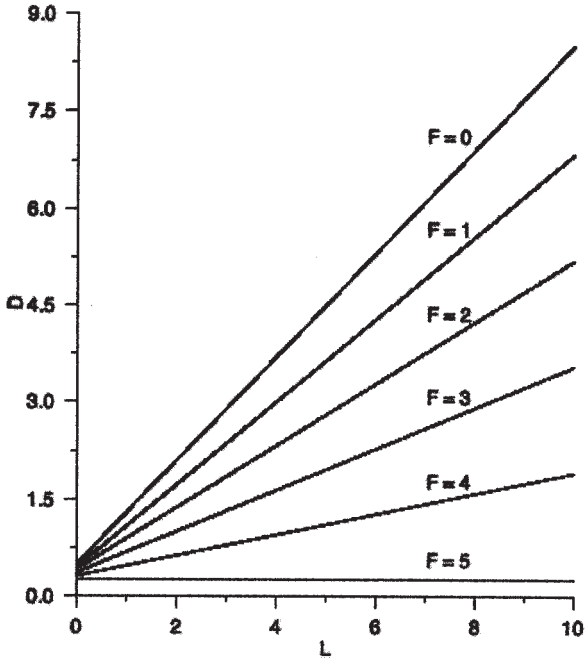

Figure 2. Depicting the moderating effect of F by plotting the simple relationship between L and D at different levels of F.

The correct interpretation of b_1 as a simple effect also explains why the coefficients for the individual components in the additive model are often so different in the moderator model; many puzzled researchers incorrectly attribute the change to the inherent collinearity between the product and its components. Collinearity is not the problem (McClelland & Judd, 1993). Instead, the coefficients in the two models are different because they are estimating very different effects. In the additive model the coefficients estimate the average or main effect, whereas in the moderator model they estimate simple effects of one variable when the other variable is fixed at 0. It is no surprise that very different questions yield very different answers.

When 0 is not a plausible value for a variable or is outside the range of the data for F in our example, the coefficient for L and its test have as little substantive interest as the intercept and its test typically have in an additive model. To avoid confusion and to focus the coefficients on more meaningful questions, it is often useful to transform the variables so that 0 becomes a meaningful value. One very useful transformation is mean deviation (or "centering"; Aiken & West, 1991; Cohen et al., 2003), in which the predictors are replaced by $L' = L - \bar{L}$ and $F' = F - \bar{F}$ so that the reexpressed model becomes $D = (b_0 + b_2F') + (b_1 + b_3F')L'$.

It is easy to show (see McClelland & Judd, 1993) that mean-deviating the predictors leaves b_3 and its test unchanged. Mean-deviating will, however, change the regression coefficients for the first-order effects of b_1 and b_2, insofar as now b_1 is the simple effect of L' on D when $F' = 0$, which is the same as the effect when F is at its average value. Cohen et al. (2003, p. 266) also noted that these simple effects of mean-deviated predictors represent "average effects of each individual predictor across the range of the other variables." This still is not quite the same as the main effect as used in traditional ANOVA terminology, but it is much closer. If the distribution of F is exactly symmetric, the estimate b_1 will be identical in both the additive and multiplicative models (i.e., identical in both steps of the hierarchical regression). Although not a necessary requirement for moderator regression analysis, we, along with Judd and McClelland (1989) and Cohen et al. (2003), strongly recommend the routine use of mean-deviated variables to reduce unnecessary confusion in interpretation of the regression coefficients (for a recent appropriate example of mean deviation, see Bradley & Corwyn, 2000).

Similarly, estimates, tests, and confidence intervals for

any simple effect of one variable at a particular level of a continuous variable are easily obtained by calculating another regression with the continuous variable deviated at the desired level. For example, to test whether the differences in family support were statistically significant at a particular level of life event stress, one could simply construct the deviated variable $L' = L - l$ and conduct another regression analysis using L', F, and $L' \times F$. Again, the interaction test will be unchanged, but the analysis will provide a test of the simple effect of F when $L = l$.[4]

One reason for not centering is if the focus is on specific values of the variables other than the mean; in that case, researchers should simply use those values to deviate their variables. For example, suppose a researcher were interested in how number of divorces (ND) moderated the relationship between two variables X and Y and particularly interested in those who had experienced exactly one divorce. Then using the variable $ND' = ND - 1$ would yield the model $Y = (b_0 + b_2ND') + (b_1 + b_3ND')X$ so that b_1 estimates the relationship between X and Y specifically for those people who have experienced exactly one divorce. The test of b_1 in this context will be much more powerful statistically than a separate analysis involving only the group experiencing exactly one divorce.

Judicious choice of codes for categorical moderators will similarly focus the tests of the individual components on interesting questions. Suppose in our running example that F is a natural dichotomy (i.e., not a split version of a multivalued variable) of family support versus no family support. Suppose the group with no family support is coded 0 and the group with family support is coded 1. Then the moderator regression model $D = b_0 + b_1L + b_2F + b_3L \times F$ can be reexpressed in terms of two regression models, one for each group:

$$D = b_0 + b_1L \text{ when } F = 0 \text{ (i.e., no family support)}$$

and

$$D = (b_0 + b_2) + (b_1 + b_3)L \text{ when } F = 1 \text{ (i.e., family support).}$$

These are in fact exactly the same models that would result if simple regressions were computed within each group, but computing them in the context of the full moderated regression model will provide much more powerful statistical tests. Note that b_0 and b_1 estimate, respectively, the intercept and the slope for the group with no family support and that b_2 and b_3 then describe how the intercept and slope differ between the no support and support groups. Most important, b_1 provides an estimate and test of the simple relationship between L and D specifically for the group with no family support. If one also desired a test of this simple relationship for those with family support, then one could simply conduct another moderated regression analysis in which the 0, 1 coding was reversed so that 0 denoted the group with family support.

No new information about the fundamental model is revealed by the second analysis; its only purpose is to provide a specific test of the relationship for those with family support. Except for sign, b_3 will be identical in the two analyses, and it tests whether the relationship between L and D differs significantly between the two groups. Bradley and Corwyn (2000) used this strategy of repeating the analysis with a change in the coding to test a two-way interaction for different ethnic groups.

Also useful is contrast coding, which can be viewed as centering a dummy code at the level of the group. Suppose the group with no family support is coded $-\frac{1}{2}$ and the group with family support is coded $\frac{1}{2}$. Then the moderator regression model $D = b_0 + b_1L + b_2F + b_3L \times F$ can be reexpressed as separate regression models for each group:

$$D = (b_0 - \tfrac{1}{2}b_2) + (b_1 - \tfrac{1}{2}b_3)L \text{ when } F = -\tfrac{1}{2}$$

and

$$D = (b_0 + \tfrac{1}{2}b_2) + (b_1 + \tfrac{1}{2}b_3)L \text{ when } F = \tfrac{1}{2}.$$

Then b_1 has the convenient interpretation as the average of the two separate within-group slopes, and b_3 is the difference between the two slopes. This focus on the average slope and the difference between two slopes is often quite useful in substantive applications of moderator regression modeling. For example, contrast coding gender produces an estimate of the average slope, regardless of whether there were equal numbers of male and female participants in the particular sample, and an estimate of the difference in the slopes between the genders. West et al. (1996) considered various coding schemes, including those described earlier, and discussed their relative advantages. They, as well as Judd and McClelland (1989), also explained how to extend coding schemes when there are three or more categorical groups.

Standardized regression coefficients. Standardized regression coefficients, or so-called "beta weights," are problematic in moderator regression models. The problem, as explained by Friedrich (1982), Jaccard and Turrisi (2003), and Aiken and West (1991), among others, is that general principles learned in the context of additive regression models do not generalize to moderator models. In the case of the additive model, most regression programs also report standardized regression coefficients ("betas") equivalent to the regression weights using standardized variables, as in $z_D = \beta_1 z_L + \beta_2 z_F$, where $z_{\cdot}$ represents the standardized form of each variable. Unfortunately, this does not generalize to moderator models including product terms because, except in very unusual circumstances, the product of two z scores

[4] Cohen et al. (2003, pp. 277–280) provided equations and Aiken and West (1991) provided computer codes for computing tests of arbitrary simple effects. However, unless there are many to be computed, most researchers will find the procedure of performing additional regression analyses with deviated variables both simpler and faster.

($z_L z_F$) will not equal the z score of the product ($z_{L \times F}$). Thus, multiple regression programs report beta or standardized regression weights equivalent to the model $z_D = \beta_1 z_L + \beta_2 z_F + \beta_3 z_{L \times F}$, but the true test and estimation of the moderator term requires using the model $z_D = \beta_0 + \beta_1 z_L + \beta_2 z_F + \beta_3 z_L z_F$.

Given these complications, it is best to avoid standardized regression coefficients entirely for moderator models and instead report raw regression weights. If standardized coefficients are required for some reason, then researchers should use proper procedures for estimating them, as described by Friedrich (1982) and Aiken and West (1991). A summary of guidelines and recommendations for appropriate analysis and interpretation of moderator regression models is provided in the Appendix.

Challenges in Testing Interactions

Having covered the basics of how to test for and interpret interactions, we shift the focus in this section to common challenges encountered in such tests for interactions. One major challenge facing researchers interested in testing interactions concerns the issue of statistical power. Although most researchers are probably familiar with the issue of statistical power for testing main effects using various analytic strategies, many investigators may be less familiar with the issue of statistical power for testing interaction effects. As previously discussed, testing the statistical significance of interactions is based on evaluating the significance of the partial correlation between the product term and the outcome when controlling for the effects of the variables included in the product term. Consequently, power calculations can be made on the basis of this partial correlation. Cohen (1988) provided a discussion of, and methods for, calculating effect sizes for this type of partial correlation. He also recommended conventions for describing small, medium, and large effect sizes for squared partial correlations of .02, .13, and .26, respectively. Using Cohen's power tables, an investigator would need to obtain sample sizes of 392, 55, and 26, respectively, to have "adequate" power (i.e., power of .80) at $\alpha = .05$ for detecting small, medium, and large effect sizes, respectively, provided that the variables included in the interaction were measured without error.

To illustrate the issue of sample size and resulting statistical power, consider the case of an investigator interested in identifying predictors of a differential response to one versus another type of treatment. For example, an investigator might be interested in whether different treatments have different effects as a function of individual differences, a paradigm that has been labeled the aptitude–treatment interaction (ATI) paradigm (Cronbach & Snow, 1977). In this case, the interaction term would be the product variable Treatment × Predictor. However, sample sizes for marital and family outcome studies are generally less than 100 and often less than 50, suggesting that most studies would have adequate power for detecting only large or perhaps, at best, medium effect sizes, again provided that the individual-differences variable was measured without error.[5]

As noted, Cohen's (1988) power tables are based on the assumption that the variables included in the interaction are measured without error. In reality, however, variables are not measured without error and therefore are not perfectly reliable. Although reliability is an issue that should always be considered in research, it becomes a particularly important issue for investigators interested in testing for interactions (Busemeyer & Jones, 1983), and thus reliability issues pose a second challenge for researchers interested in testing interaction effects. As discussed by Cohen et al. (2003), the reliability of the product term of two mean-deviated variables with uncorrelated true scores is the product of the reliabilities of the two variables. For example, if two variables each have a reliability of .80, then the reliability of the interaction term will be .64 (.8 × .8). When X and Z are uncorrelated, these two variables must have average reliabilities of .89 and .84 for the interaction term to have a reliability of .80 and a reliability of .70, respectively (Aiken & West, 1991). Insofar as the effect of unreliability is that it reduces the association between variables, "when individual predictors are less than perfectly reliable, the interaction term is even more unreliable, and we expect the power to detect the interaction term to be reduced, relative to the power to detect the first-order effects, even if they have equal effect sizes in the population" (Cohen et al., 2003, p. 297). Consequently, the required sample size for having power to detect interactions will be larger, with the size of the increase varying as a function of the reliability of the measures included in the product term. As discussed in greater detail by Aiken and West (1991), the sample size required to reach a power of .80 with an alpha value of .05 is slightly more than doubled when reliabilities drop from 1.0 to .80 and more than tripled when reliabilities drop from 1.0 to .70. Furthermore, required sample sizes will be even larger in situations in which the variables included in the interaction are themselves associated with the outcome variable: "The greater the proportion of variance accounted for by the first order effects, the sharper is the decline in the effect sizes, variance accounted for, and power of the test for the interaction term as reliability decreases" (Aiken & West, 1991, p. 163). In summary, samples of more than 200 participants may be necessary for having adequate power for detecting interactions with medium effect sizes using measures with reliabilities of .70, whereas more than 1,000 participants may be necessary for detecting interactions with small effect sizes. Therefore, failure to find hypothesized interactions in family research may be due, in many cases, to insufficient sample sizes and resulting low statistical power.

A third challenge facing researchers interested in testing

[5] It should be noted that the difficulty in detecting predictors of differential response to one versus another treatment should not be construed to mean that an investigator may not have adequate power to detect predictors of response to a particular treatment or response across treatments, which can be tested, for example, by computing the association between a predictor and posttreatment outcome controlling for pretreatment scores on the outcome variable.

interactions concerns the distribution of variables included in the interaction. McClelland and Judd (1993) discussed several implications of the joint distribution of the two variables included in the product term in regard to detecting interactions. First, as is well known, effect sizes are determined, in part, by the variability of the measures, with variables with restricted ranges resulting in smaller effect sizes. As was the case with reliability, McClelland and Judd demonstrated that problems with restricted range are exacerbated (or compounded) when testing for interaction effects. The effects of reduced variance are exactly analogous to the effects of reduced reliability. If, for whatever reasons, the variances of the two predictors were reduced to 80% of what they otherwise would have been, then the variance of their product—the interaction term—would only be .64 (.8 × .8) of what it otherwise would have been. In particular, problems with restricted range result in few participants having extreme values on both variables included in the interaction. To use our earlier example, problems with restricted range would suggest that only a small percentage of participants would have very low (or high) levels of family support *and* very low (or high) levels of life stress. Other problems such as clustering of cases in the center of the distribution (i.e., a normal distribution) or distribution of observations over many categories also serve to reduce the variance in the individual predictor variables.

McClelland and Judd (1993) then discussed the efficiency of detecting interactions with various joint distributions of two variables. They concluded that "jointly extreme observations are crucial for detecting interactions" (McClelland & Judd, 1993, pp. 382–383). Thus, the optimal design for detecting interactions is a "four-corners" design in which 25% of cases are allocated to each extreme (i.e., 25% of cases are extremely high on both variables, 25% are extremely low on both variables, and 25% are extremely high on one variable and are extremely low on the other variable).[6] In comparison with the four-corners design, a normal-like distribution of the two variables has a relative efficiency of only .06 for detecting an interaction and requires nearly 17 times as many observations to have comparable relative efficiency (McClelland, 1997). Unfortunately, however, field studies of the kind often used in family research result in a proportionately small number of extreme cases, leading to nonoptimal distributions of variables and, consequently, less powerful tests of interaction effects.

Aware of the impact that the joint distribution of variables has on detecting interaction effects, investigators might be tempted to use median splits to recode their variables into fewer categories. However, as has already been discussed, using median splits results in loss of information and reduced statistical power. Similarly, other potential solutions such as using multiple cut points or selecting measures with fewer categories result in the variables of interest being measured with less precision (i.e., more error), and thus, they should be avoided. As discussed in the next section, the solution to this problem lies in obtaining a sufficient number of participants with truly extreme scores on the measures included in the interaction.

A fourth important challenge to detecting interactions in family research is that generally only ordinal interactions (i.e., slopes have the same sign over the range of the variables) are theoretically expected. For example, although family support is likely to ameliorate the negative effects of life event stress and perhaps even eliminate those effects, it is unlikely to convert life event stress into positive effects. Similarly, the ATI for a treatment versus control study is likely to be differential effectiveness of the treatment as a function of the predictor but no effectiveness of the control condition regardless of the level of the predictor, that is, a sloped regression line for the treatment group and a flat regression line for the control group. In the case of ordinal dichotomous by dichotomous interactions (i.e., an interaction in a 2 × 2 ANOVA), Rogers (2002) proved that the increment in the squared correlation due to the interaction is severely constrained by the squared correlation of the additive model; in particular,

$$\Delta R^2_{Mod} \leq \frac{R^2_{Add}}{2}.$$

This is a profound constraint. One implication is that the maximum possible value of ΔR^2_{Mod} is .33, and that maximum could be achieved only in the virtually impossible case in which the overall R^2 value was 1.0. In a more realistic case in which, say, $R^2_{Add} = .4$, the maximum possible value for ΔR^2_{Mod} is only .2. The constraints for dichotomous by continuous interactions are more severe, and those for continuous by continuous interactions are still more severe. The exact constraints are too complex to describe here. On the basis of his analysis of those constraints, Rogers (2002, p. 221) concluded that, for interactions of continuous variables that have nonnegative correlations, "the existence of 'medium' [partial squared correlation = .13] interaction effect sizes will only occur at levels of R^2_{Add} of .50 or greater. [An] R^2_{Add} of .30 will only allow ordinal interactions of effect sizes up to approximately .06 when [the predictor intercorrelation] is near zero."[7] Not surprisingly, Champoux and Peters (1987) and Chaplin (1991) found, in field studies in the social sciences, that ΔR^2_{Mod} typically ranges only from .01 to .03. This suggests that the sample size estimates described earlier for Cohen's (1988) small, medium, and large effect sizes, discouraging as they were, are much too optimistic for ordinal interactions. A useful summary of Rogers's constraints is that if one is to be able to detect an ordinal interaction, there must be a strong effect (i.e., a strong association between *L* and *D*) to be moderated for at least some level of the moderator variable (for continuously

[6] The four-corners design can detect only linear by linear interactions. McClelland (1997) described how to add a fifth middle group to test for the existence of higher order interactions without substantially decreasing power for detecting the most likely linear by linear interaction.

[7] We have replaced Rogers's (2002) notation and effect size index to be consistent with those used elsewhere in this article.

distributed moderators) or for one group (for dichotomous moderators).

A frequently cited challenge to the detection of interactions that is in fact not a problem is multicollinearity. This is another example of intuitions developed for additive multiple regression that do not generalize to moderator multiple regression. Commonly, the product $X \times Z$ is highly correlated with the components X and Z, thereby creating an apparent multicollinearity problem. However, we know that (a) changing the origin of the scales, for example, by centering, does not affect the test of the interaction, and (b) a change of origin always exists so as to reduce the correlation between $X' \times Z'$ and its components to zero (Aiken & West, 1991; Friedrich, 1982; McClelland & Judd, 1993). Not only is this aspect of multicollinearity a non-issue, but McClelland and Judd (1993) and Rogers (2002) proved that correlation between the two predictors X and Z slightly facilitates the detection of interactions by producing more cases that are jointly extreme.

Finally, in discussing challenges in evaluating interactions, we should note that researchers are well advised to use regression diagnostics to check for outliers and violations of assumptions when conducting any regression analysis, but such checks are even more important in moderated regression analysis because outliers and assumption violations can be more problematic. Busemeyer and Jones (1983) showed that scaling issues of the kind likely to be associated with violations of the homogeneity of variance assumption are especially problematic for moderator analyses. Lubinski and Humphreys (1990) demonstrated that nonlinearity of the predictor variables can produce spurious interactions. Judd and McClelland (1989) presented an example in which the interaction disappeared after appropriate attention to outliers and transformation to achieve homogeneity of variance, whereas McClelland (2000), in a chapter describing a number of useful regression diagnostics, provided an example in which a clear interaction emerged only after attention to outliers and transformation to achieve homogeneity of variance. The message is that detection of meaningful interactions is very sensitive to modeling and statistical assumptions.

Recommendations for Designing Research to Test for Interactions

In the preceding section, we reviewed design issues that pose challenges for investigators who are interested in detecting interactions in family research. In this section, we offer recommendations for designing studies to test for interactions that address these challenges.

First, insofar as the statistical power for detecting interactions is greatly influenced by the reliability of the measures used in the interaction, one obvious implication for designing future studies would be to select measures with good reliability. This may seem like a trite recommendation, insofar as psychometric properties of measures should always be considered in selecting measures for any research study. However, whereas an investigator may be willing to use a measure with borderline acceptable reliability (i.e., reliability of .70) for studies evaluating bivariate associations, such a measure may not be acceptable for studies involved in testing interactions. Therefore, greater attention to issues of reliability of measures is warranted in designing studies for detecting interactions.[8] In a related fashion, investigators who already have data based on measures of lower reliability may want to reconsider using these data in tests of interactions, unless they have a very large sample, because they are unlikely to have adequate statistical power for detecting interaction effects.

Second, in addition to selecting measures with good reliability, incorporating other methods for increasing statistical power should also be considered in testing interactions. For example, investigators may want to consider adopting a higher alpha level (which may be unacceptable to journal editors) or increasing the sample size (McClelland & Judd, 1993).

Third, insofar as tests of interactions are greatly influenced by the joint distribution of the variables used in the interaction, an important implication for designing future studies would be to consider issues of sampling individuals for participation. Rather than randomly sampling people for participation, investigators may want to consider oversampling people with extreme scores. Such a practice would result in a larger number of people with extreme values on both predictor variables, which would make it easier to detect interaction effects. To use our ongoing example, this would entail oversampling participants who scored very high or very low on family support or life events. As noted by McClelland and Judd (1993), this strategy is controversial insofar as the overall variance explained (i.e., R^2) by the model will be inflated. However, if investigators are aware of this, then they can appropriately discuss their findings in terms of detecting whether or not an interaction exists and not in terms of the magnitude of such an interaction effect. For example, if the goal of the study is not to estimate the population effect size but to identify interactions that would be potentially useful in designing a treatment intervention, then a researcher might want to oversample extreme (or "abnormal") cases and undersample middle (or "normal") cases for whom the treatment intervention would not be appropriate at any rate.

Fourth, because the magnitude of the moderator effect for ordinal interactions is constrained by the additive effect, researchers should expect to find interactions only for variables with strong effects for at least some levels of the moderating variable. For example, if a treatment has a weak effect even in the best of circumstances, then it will be virtually impossible to identify conditions in which that weak effect is even weaker or nonexistent. The search for interactions begins with a search for strong variables and potent treatments.

[8] Kenny and Judd (1984) also showed that these problems can be ameliorated through structural equation modeling.

Conclusion

We have provided a primer for testing and interpreting interactions, including discussing common pitfalls to avoid in this kind of research. The only appropriate statistical test of interactions is a comparison of the additive model and the moderator model in which the product or products of the additive components have been added. We also have provided careful advice on how to interpret regression coefficients in moderator models, an issue about which there is much confusion in published results. In addition, we have discussed several difficulties in detecting interactions, including issues related to statistical power, measurement error, distribution of variables, and mathematical constraints of ordinal interactions. In general, the usual problems plaguing research are even more problematic when testing for interactions. Finally, we have provided several recommendations for improving the design of studies evaluating interactions, including using reliable measures, enhancing statistical power, oversampling extreme scores, and using strong variables and potent treatments. Adherence to the guidelines and recommendations covered in this article should assist readers in understanding existing moderator research as well as improve future research on interaction effects in family psychology.

References

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

Bradley, R. H., & Corwyn, R. F. (2000). Moderating effect of perceived amount of family conflict on the relation between home environmental processes and the well-being of adolescents. *Journal of Family Psychology, 14,* 349–364.

Busemeyer, J. R., & Jones, L. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93,* 549–562.

Champoux, J. E., & Peters, W. S. (1987). Form, effect size, and power in moderated regression. *Journal of Occupational Psychology, 60,* 243–255.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59,* 143–178.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

Cohen, J. (1978). Partialed products *are* interactions; partialed vectors *are* curve components. *Psychological Bulletin, 85,* 858–866.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26,* 797–833.

Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research, 38,* 100–109.

Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research, 40,* 366–371.

Jaccard, J. (2001). *Interaction effects in logistic regression.* Thousand Oaks, CA: Sage.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression* (2nd ed.). Thousand Oaks, CA: Sage.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, pp. 180–232). New York: McGraw-Hill.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 90,* 201–210.

Liew, J., Eisenberg, N., Losoya, S. H., Fabes, R. A., Guthrie, I. K., & Murphy, B. C. (2003). Children's physiological indices of empathy and their socioemotional adjustment: Does caregivers' expressivity matter? *Journal of Family Psychology, 17,* 584–597.

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin, 107,* 383–393.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology, 46,* 1–26.

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.

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

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

McClelland, G. H. (2000). Nasty, ill-mannered observations can ruin your analysis. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 393–411). Cambridge, England: Cambridge University Press.

McClelland, G. H., Irwin, J. R., & Judd, C. M. (2004). *Continuous to discrete transformations: Decreased statistical power and increased bias.* Manuscript in preparation.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

Rogers, W. M. (2002). Theoretical and mathematical constraints of interactive regression models. *Organizational Research Methods, 5,* 212–230.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics, 21,* 264–282.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality, 64,* 1–48.

(*Appendix follows*)

Appendix

Summary of Guidelines for Testing and Interpreting Interactions

1. Estimate the moderator effect as b_3 in the moderator regression model $D = b_0 + b_1L + b_2F + b_3L \times F$ and test the interaction as the increment in the squared correlation between the additive and moderator models.

2. If the product $L \times F$ is included in the regression model, then both components of the product (L and F) must be included in the model.

3. Visually represent and interpret a moderator regression model by plotting one of the two independent variables on the x-axis and then using interesting values of the other independent variable to define lines relating the x-axis variable to the dependent variable.

4. Use mean deviation or other changes of origin and coding of categorical variables to create predictors for which 0 is a meaningful value (i.e., any particular level of the other variable). The coefficients of the individual terms that are included in products then represent the relationship of that variable to the dependent variable when the *other* independent variable is fixed at 0.

5. Use the following prototypes for interpreting the "raw" regression coefficients. For coefficients of individual terms: "When F is fixed at 0, a one-unit increase in L predicts an increase of b_1 in the dependent variable D." For the moderator coefficient: "For each one-unit increase in the moderator F, the slope of the relationship between L and D increases by b_3."

6. Do not use (a) a test of differences among subgroup correlations as a test of moderation or interaction, (b) median splits or other dichotomizations of multivalue variables for any reason, or (c) standardized regression weights reported by multiple regression programs.

Received January 15, 2004
Revision received June 1, 2004
Accepted September 28, 2004 ■