*Multivariate Behavioral Research*, *39* (1), 99-128


# Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods

David P. MacKinnon, Chondra M. Lockwood, and Jason Williams
Arizona State University

The most commonly used method to test an indirect effect is to divide the estimate of the indirect effect by its standard error and compare the resulting *z* statistic with a critical value from the standard normal distribution. Confidence limits for the indirect effect are also typically based on critical values from the standard normal distribution. This article uses a simulation study to demonstrate that confidence limits are imbalanced because the distribution of the indirect effect is normal only in special cases. Two alternatives for improving the performance of confidence limits for the indirect effect are evaluated: (a) a method based on the distribution of the product of two normal random variables, and (b) resampling methods. In Study 1, confidence limits based on the distribution of the product are more accurate than methods based on an assumed normal distribution but confidence limits are still imbalanced. Study 2 demonstrates that more accurate confidence limits are obtained using resampling methods, with the bias-corrected bootstrap the best method overall.

An indirect effect implies a causal hypothesis whereby an independent variable causes a mediating variable which, in turn, causes a dependent variable (Sobel, 1990). Hypotheses regarding indirect or mediated effects are implicit in social science theories (Alwin & Hauser, 1975; Baron & Kenny, 1986; Hyman, 1955; Sobel, 1982). Examples of indirect effect hypotheses are that attitudes affect intentions which then affect behavior (Ajzen & Fishbein, 1980), that poverty reduces local social ties which increases assault and burglary rates (Warner & Rountree, 1997), that social status has an indirect effect on depression through changes in social stress (Turner, Wheaton, & Lloyd, 1995), and that father's education affects offspring education which then affects offspring income (Duncan, Featherman, & Duncan, 1972).

---

This research was supported by the National Institute on Drug Abuse grant number 1 R01 DA09757. We acknowledge the contributions of Ghulam Warsi and Jeanne Hoffman to the work described in this article. We thank William Meeker, Leona Aiken, Michael Sobel, Steve West, and Jenn-Yun Tein for comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to David P. MacKinnon, Department of Psychology, Arizona State University, Tempe, AZ 85287-1104.

Analysis of indirect effects is also important for experimental studies of social policy interventions. Substance abuse prevention programs, for example, are designed to change mediating variables such as social bonding (Hawkins, Catalano, & Miller, 1992) and social influence (Bandura, 1977) which are hypothesized to be causally related to drug abuse (see also Hansen & Graham, 1991, and Tobler, 1986, for more examples). In these contexts, the randomization of participants to treatment conditions and the knowledge that the treatment precedes both the mediating variable and the dependent variable in time strengthen the causal inferences that may be drawn about the indirect effects of the intervention (Holland, 1988; Sobel, 1998). In these experimental studies, analysis of indirect effects (also called mediation analysis) provides a check on whether the manipulation changed the variables it was designed to change, tests theory by providing information on the process through which the experiment changed the dependent variable, and generates information that may improve programs (MacKinnon, 1994; West & Aiken, 1997). Thus, the accuracy of confidence limits for the indirect effect is important for both basic and applied researchers in several substantive areas of social science (Allison, 1995a; Bollen & Stine, 1990; Sobel, 1982).

Prior research provides much information on the relative performance of various methods for conducting significance tests for indirect effects (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002), but very little information about confidence limits. Confidence limit estimation has been advocated for several reasons, including that it forces researchers to consider the size of an effect in addition to making a binary decision regarding significance and that the width of the interval provides a clearer understanding of variability in the size of the effect (Harlow, Mulaik, & Steiger, 1997; Krantz, 1999). The purpose of this article is to explain why the traditional method used to test the significance of the indirect effect based on the assumption of the $z$ distribution has statistical power and Type I error rates that are too low and imbalanced confidence limits. Two alternatives to address the problem are evaluated in this article, one based on the distribution of the product of two normal random variables and another based on resampling methods. First, the equations used to estimate the indirect effect and its standard error are described, followed by evidence that traditional confidence limits for the indirect effect are imbalanced. Next, an overview of the distribution of the product is given with a description of how this distribution explains inaccuracies in the traditional test of the indirect effect. In Study 1, confidence limits for the traditional and distribution of the product methods are compared in a statistical simulation. In Study 2, a simulation study compares the distribution of the product method evaluated

in Study 1 with resampling methods which should also adjust for the nonnormal distribution of the indirect effect.

*Estimation of the Indirect Effect and Standard Error*

The indirect effect model is shown in Figure 1 and is summarized in the three equations described below (see also Allison, 1995a and MacKinnon & Dwyer, 1993). We focus on a recursive model with a single indirect effect and ordinary regression models in order to more clearly describe the approach.

$$Y_O = \beta_{01} + \tau X + \varepsilon_1 \quad (1)$$

$$Y_O = \beta_{02} + \tau' X + \beta X_M + \varepsilon_2 \quad (2)$$

$$X_M = \beta_{03} + \alpha X + \varepsilon_3 \quad (3)$$

In these equations, $Y_O$ is the dependent variable, $X$ is the independent variable, $X_M$ is the mediating variable, $\tau$ codes the relation between the independent variable and the dependent variable, $\tau'$ codes the relation between the independent variable and the dependent variable adjusted for the effects of the mediating variable, $\alpha$ codes the relation between the independent variable and the mediating variable, and $\beta$ codes the relation between the mediating variable and the dependent variable adjusted for the independent variable. The residuals are coded by $\varepsilon_1$, $\varepsilon_2$, and $\varepsilon_3$ and the intercepts are coded by $\beta_{01}$, $\beta_{02}$, and $\beta_{03}$ in Equations 1, 2, and 3, respectively. The residuals have expected values of zero.

In the first regression equation, the dependent variable ($Y_O$) is regressed on only the independent variable ($X$). In the second regression equation, the dependent variable ($Y_O$) is regressed on both the independent variable ($X$) and the mediating variable ($X_M$). The indirect effect equals the difference in the estimated independent variable coefficients ($\hat{\tau} - \hat{\tau}'$) in the two regression equations (Judd & Kenny, 1981).

A second method to calculate the indirect effect is illustrated in Figure 1. First, the coefficient relating the mediating variable to the dependent variable is estimated ($\hat{\beta}$) in Equation 2 above. Second, the coefficient relating the independent variable to the mediating variable is estimated ($\hat{\alpha}$) in Equation 3. The product of these two estimates ($\hat{\alpha}\,\hat{\beta}$) is the estimated indirect effect. The estimated coefficient relating the independent variable to the dependent variable adjusted for the mediating variable ($\hat{\tau}'$) is the estimate of the direct effect. The $\hat{\tau} - \hat{\tau}'$ and $\hat{\alpha}\,\hat{\beta}$ estimators of the indirect effect are equivalent

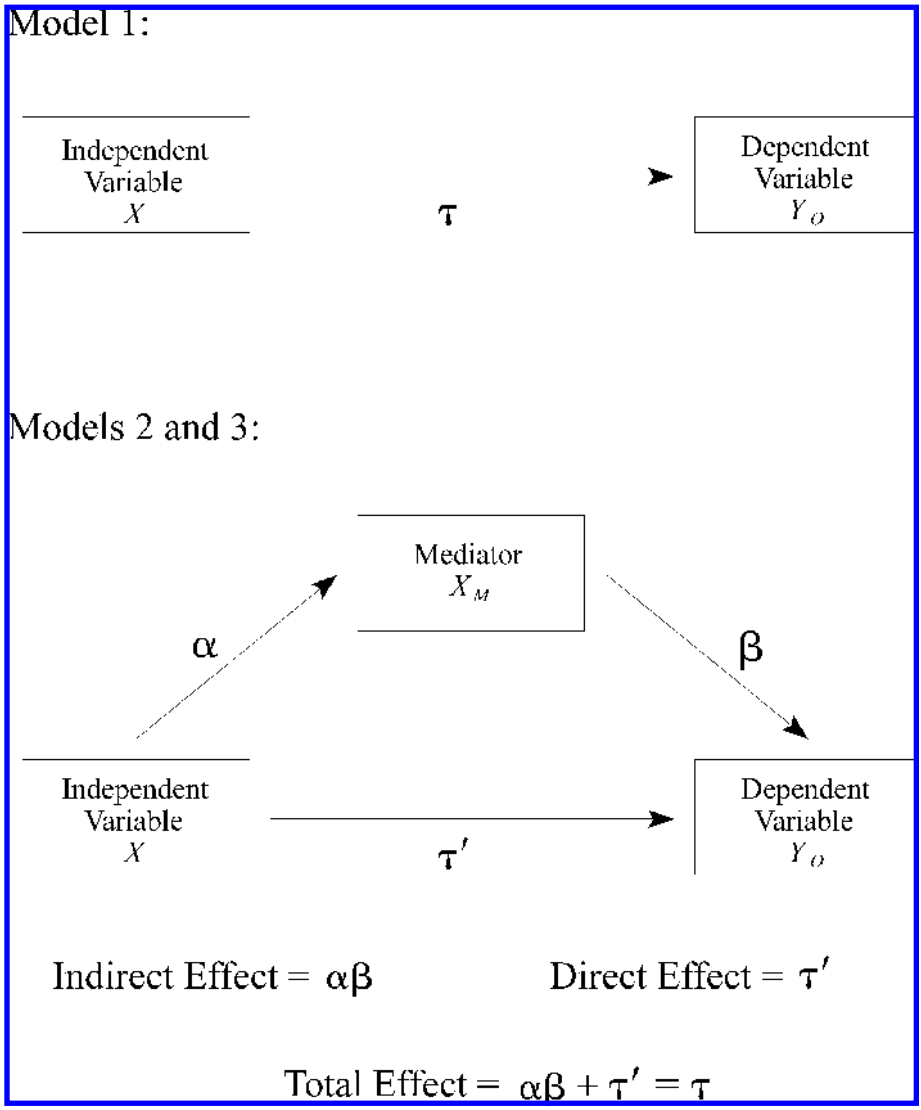


**Figure 1**
The Indirect Effect Model

in ordinary least squares regression (MacKinnon, Warsi, & Dwyer, 1995). Additional assumptions of the $\hat{\alpha}\,\hat{\beta}$ estimator of the indirect effect from Equations 2 and 3 have been outlined (James & Brett, 1984; McDonald, 1997). These assumptions include no measurement error in variables (Hoyle & Kenny, 1999), the causal relations of *X* to *M* to *Y* are correct (McDonald, 1997), no omitted variables (McDonald, 1997), and a zero interaction of *X* and $X_M$ (Judd & Kenny, 1981). The same assumptions are made for the indirect effect model examined in this article.

Although there are several estimators of the variance of the indirect effect (see MacKinnon et al., 2002), the most commonly used estimator was derived by Sobel (1982; 1986). This formula (Equation 4), based on the multivariate delta method, is used to calculate the standard error of the indirect effect in statistical software packages, including EQS (Bentler, 1997), LISREL (Jöreskog & Sörbom, 1993), and LINCS (Schoenberg & Arminger, 1996), and is based on the estimates $\hat{\alpha}$ and $\hat{\beta}$, and the estimated standard errors, $\hat{\sigma}_{\hat{\alpha}}$ and $\hat{\sigma}_{\hat{\beta}}$. Allison (1995a) used a reduced form parameterization of the indirect effect model to derive the same standard error formula in Equation 4. The formula assumes that $\alpha$ and $\beta$ are independent (Sobel, 1987). This variance estimator can be used to calculate standard errors and confidence limits for the indirect effect. MacKinnon and Dwyer (1993) and MacKinnon et al. (1995, 2002) found evidence that the multivariate delta method standard error had the least bias of several formulas for the standard error of the indirect effect.

$$\hat{\sigma}^2_{\hat{\alpha}\hat{\beta}} = \hat{\alpha}^2\hat{\sigma}^2_{\hat{\beta}} + \hat{\beta}^2\hat{\sigma}^2_{\hat{\alpha}} \tag{4}$$

For nonzero values of both $\alpha$ and $\beta$, simulation studies suggest that the variance estimator has relative bias less than 5% for sample sizes of 100 or more in a single indirect effect model (MacKinnon et al., 1995) and for sample sizes of 200 or more in a recursive model with seven total indirect effects (Stone & Sobel, 1990).

In many studies, the indirect effect is divided by its standard error and the resulting ratio is then compared to the standard normal distribution to test its significance, $z = \hat{\alpha}\hat{\beta}/\hat{\sigma}_{\hat{\alpha}\hat{\beta}}$ (Bollen & Stine, 1990; MacKinnon et al., 1991; Wolchik, Ruehlman, Braver, & Sandler, 1989). Confidence limits for the indirect effect lead to the same conclusion with regard to the null hypothesis. Confidence limits are constructed using Equation 5,

$$\hat{\alpha}\hat{\beta} \pm z_{1-\omega/2} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}} \tag{5}$$

where $z_{1\,-\,\omega/2}$ is the value on the standard normal distribution corresponding to the desired Type I error rate, $\omega$.

Although the variance and standard error estimators of the indirect effect may be unbiased at small sample sizes, there is evidence that confidence limits based on these values do not perform well. Two extensive simulation studies (MacKinnon et al., 1995; Stone & Sobel, 1990) showed an imbalance in the number of times a true value fell to the left or right of the confidence limits. For positive values of the indirect effect, where $\alpha$ and $\beta$ are both positive or both negative, the true value was more often to the right than to the left of the confidence interval. Asymmetric confidence intervals were also obtained in bootstrap analysis of the indirect effect (Bollen & Stine, 1990; Lockwood & MacKinnon, 1998). The implication of the imbalance is that there is less power than expected to detect a true indirect effect. Stone and Sobel (1990, p. 349) analytically demonstrated that at least part of the imbalance may be due to use of only first order derivatives in the solution for the multivariate delta standard error of the indirect effect. However, MacKinnon et al. (1995) found that the second order Taylor series solution had similar imbalances in confidence intervals. Bollen and Stine (1990) noted that the asymptotic solutions for the standard error of the indirect effect may be incorrect for small sample sizes and used a bootstrapping approach to improve the accuracy of confidence limits for the indirect effect. An explanation for the low power and imbalance in confidence limits is the assumption that the distribution of the indirect effect is normal when, in fact, it is skewed for nonzero indirect effects and has different values of kurtosis for different values of the indirect effect, as will be described in the next section.

*The Distribution of the Product*

The assumption that the indirect effect divided by its standard error has a normal sampling distribution is incorrect in some situations. In these situations, the confidence limits calculated using Equation 5 will be incorrect. Because the indirect effect is the product of regression estimates which are normally distributed asymptotically (Hanushek & Jackson, 1977), an alternative method for testing indirect effects can be developed based on the distribution of the product of two normally distributed random variables (Aroian, 1947; Craig, 1936; Springer, 1979).

The product of two normal random variables is not normally distributed (Lomnicki, 1967; Springer & Thompson, 1966). In the null case where both random variables have means equal to zero, the distribution is symmetric with kurtosis of six (Craig, 1936). When the product of the means of the two random variables is nonzero, the distribution is skewed as well as having excess kurtosis, although Aroian (1947) and Aroian, Taneja, and Cornwell

(1978) showed that the product approaches the normal distribution as one or both of the ratios of the means to standard errors of each random variable get large in absolute value. The four moments of the product of two correlated normal variables are given in Craig (1936), Aroian et al. (1978), and Meeker, Cornwell, and Aroian (1981).

The general analytical solution for the distribution of the product of two independent standard normal variables does not approximate distributions commonly used in statistics, although Aroian (1947) showed that the gamma distribution can provide an approximation in some situations. Instead, the analytical solution for this product distribution is a Bessel function of the second kind with a purely imaginary argument (Aroian, 1947; Craig, 1936). Although computation of these values is complex, Springer and Thompson (1966) provided a table of the values of this function when $\alpha = \beta = 0$. Meeker et al. (1981; see pages 129-144 for uncorrelated variables) presented tables of the distribution of the product of two normal random variables based on an alternative formula more conducive to numerical integration. Tables of fractiles of the standardized distribution function for

$$\frac{\hat{\alpha}\hat{\beta} - \alpha\beta}{\sigma_{\hat{\alpha}\hat{\beta}}}$$

for different values of $\alpha$, $\beta$, $\sigma_\alpha$, and $\sigma_\beta$ were given in Meeker et al. (1981). The denominator was the exact standard error (Aroian, 1947; MacKinnon et al., 2002). Table entries were in terms of the ratio of each of the two variables to their standard error so for $\alpha$ and $\beta$, the table entries corresponded to $\delta_\alpha = \alpha/\sigma_\alpha$ and $\delta_\beta = \beta/\sigma_\beta$. Note that sample values of $\delta_\alpha$ and $\delta_\beta$ correspond to the *t*-values for each regression coefficient. The critical values in the table assume that population values of $\alpha$, $\beta$, $\sigma_\alpha$, and $\sigma_\beta$ are known, but the authors suggest that sample values can be used in place of the population values as an approximation (Meeker et al., 1981, p. 8).

The frequency distribution function of the product of two standardized normal variables, presented in Meeker et al. (1981), can be used to compute the statistical power, Type I error, and confidence limits for the indirect effect. The test based on the tables in Meeker et al. is represented by *M* in this article. In this method, the two estimated $\delta$ statistics, one for the $\hat{\alpha}$ estimate, $\hat{\delta}_\alpha$, and another for the $\hat{\beta}$ estimate, $\hat{\delta}_\beta$, are computed and used to look up critical values in the table of critical values in Meeker et al. (1981). For example, the critical values of the *M* distribution for $\delta_\alpha = 0$ and $\delta_\beta = 0$ (a symmetric distribution) for two-tailed tests are 3.60, 2.18, and 1.60 for nominal Type I error rates of 0.01, 0.05, and 0.10, respectively.

The confidence limits based on the distribution of the product using Equation 5 require different critical values for the upper (UCL) and lower (LCL) confidence limits because the distribution of the product of two normal random variables can be asymmetric and depends on the values of $\delta_\alpha$ and $\delta_\beta$. The critical values are read from tables in Meeker et al. (1981) based on values of $\hat{\delta}_\alpha$ and $\hat{\delta}_\beta$ and Equations 6 and 7 are used to compute the upper and lower confidence limits, respectively. This test is called the asymmetric distribution of the product test in MacKinnon et al. (2002).

(6) $$UCL = \hat{\alpha}\,\hat{\beta} + \text{Meeker Upper} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}}$$

(7) $$LCL = \hat{\alpha}\,\hat{\beta} + \text{Meeker Lower} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}}$$

The critical values are obtained from the tables in Meeker et al. (1981) where Meeker Upper is the critical value for the upper confidence limit and Meeker Lower is the critical value for the lower confidence limit. For example, the 95% confidence interval for $\delta_\alpha = .4$ and $\delta_\beta = 1.2$ would be calculated using a Meeker Upper critical value of 2.3774 (p. 141) and a Meeker Lower critical value of -1.8801 (p. 131).

## *Study 1*

As described above, there is evidence that the traditional *z* test of the indirect effect has imbalanced confidence limits. The purpose of Study 1 is to compare the traditional *z* test to an approach based on the distribution of the product of two normal random variables in a large statistical simulation study.

### *Methods*

#### *Simulation Description*

The SAS® (1989) programming language was used to conduct the statistical simulations. The data were simulated using Equations 2 and 3, with sample values of *X*, $\varepsilon_2$, and $\varepsilon_3$ generated from a standard normal distribution using the SAS RANNOR function with current time as the seed for each simulation. Five different sample sizes corresponding to sample sizes in the social sciences were simulated: 50, 100, 200, 500, and 1000. The independent variable was simulated to be a normally distributed continuous variable. A simulation study with a binary independent variable led to the same results as for the continuous independent variable case, so they are not described in

this article. All combinations of four parameter values of $\alpha$ and $\beta$ were simulated: 0, .14, .39, and .59. The different effect sizes corresponded to zero, small (2% of the variance), medium (13% of the variance), and large (26% of the variance) effect sizes as described in Cohen (1988, p. 412-414). Previous simulations indicated no difference in power calculations as the direct effect ($\tau'$) increased, so for simplicity the direct effect was always set equal to zero.

The four effect sizes (zero, small, medium, or large) for $\alpha$ and $\beta$, and five sample sizes (50, 100, 200, 500, or 1000) yielded a total of 80 different combinations. Ten thousand replications were conducted for each of the 80 combinations.

*Confidence Limits*

Sample values were inserted in Equation 5 to obtain upper and lower confidence limits for the *z* test. The value of 1.96 was used for $z_{1 - \omega/2}$ and the standard error was the multivariate delta solution in Equation 4.

Equations 6 and 7 were used to calculate the *M* confidence limits. The upper and lower critical values were obtained from the table in Meeker et al. (1981) for percentiles of .025 and .975. The critical values for the confidence limits were selected based on the sample estimates, $\hat{\delta}_\alpha$ and $\hat{\delta}_\beta$, and were used with the standard error from Equation 4. Sample values, $\hat{\delta}_\alpha$ and $\hat{\delta}_\beta$, were used because researchers will typically not know the true values of $\delta_\alpha$ and $\delta_\beta$.

The accuracy of confidence limits was assessed by calculating the proportion of times the true value was above the upper confidence limit and the proportion of times the true value was below the lower confidence limit and comparing the proportion to the predicted proportion. For example, for 95% confidence intervals, the most accurate methods should have the proportion of true values below the lower limit equal to .025 and the proportion of true values above the upper confidence limit equal to .025. The liberal robustness criterion given by Bradley (1978) was used to assess whether or not the proportions deviated significantly from the expected proportion. Proportions were considered robust if they fell within the range .5$\omega$/2 to 1.5$\omega$/2.

*Results*

Confidence limits were calculated for the *z* and *M* tests. The proportions of times that true values of the indirect effect fell to the left and right of the confidence limits are shown in Tables 1 and 2. Because results for certain

parameter combinations were virtually identical, only nonredundant combinations are reported in the table (e.g., $\alpha = 0$ and $\beta = .14$ but not $\alpha = .14$ and $\beta = 0$). As found in other studies of the confidence limits of the indirect effect assuming normal sampling distributions, there is an imbalance in the *z* test confidence limits. The distribution of the product, *M*, has more balanced confidence limits because it incorporates the skewness and kurtosis of the product distribution. Most importantly, note that the confidence limits based on the distribution of the product are nearly always as close or closer to the expected Type I error rate of .025 than the traditional test. There were two exceptions to this pattern in the 80 possible intervals investigated in the simulation study.

For both methods of computing the confidence interval, the nominal Type I error rates were often different from the expected values (.025 outside the upper and lower confidence limits) when one or both parameters had a zero or small effect size. As sample size increased, the accuracy of both methods improved, consistent with the tendency for the *M* distribution to approach the normal distribution as sample size increases when either $\alpha$ or $\beta$ or both are nonzero. In fact, the critical values based on the distribution of the product test are the same as the critical values from the normal distribution as either or both $\delta_\alpha$ and $\delta_\beta$ get very large.

### *Discussion*

The confidence limits for the indirect effect based on the distribution of the product were more accurate than the confidence limits based on the normal distribution assumption. The Type I error rates based on the confidence limits did not exceed nominal rates using this method for any combination of parameter values both in Study 1[1] and in MacKinnon et al. (2002). Despite being more accurate than confidence limits constructed with the normal distribution, the proportions outside the confidence limits for the product distribution were often less than the expected values for small effect sizes and small sample sizes (i.e., small values of $\delta_\alpha$ and $\delta_\beta$). One possible explanation for this discrepancy is that the appropriate comparison distribution is the product of two *t* distributions rather than two normal distributions. The distribution of the product of two variables with *t* distributions may be more complicated than the distribution of the product of two normal variables.

[1] Although not reported in this article because of space limitations, we compared expected and empirical power and Type I error rates for the traditional *z* test and the distribution of the product test. The expected and empirical power and Type I error rates were very close for the distribution of the product test but were discrepant for the traditional *z* test for smaller sample sizes and effect sizes. More on this topic is at our website http://www.public.asu.edu/~davidpm/.

Table 1
Proportion of True Values to Left and Right of 95% Confidence Intervals - One or Two Zero Paths, Study 1

| Effect Size | | | Sample Size | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 50 | | 100 | | 200 | | 500 | | 1000 | |
| α | β | Test | true to left | true to right | true to left | true to right | true to left | true to right | true to left | true to right | true to left | true to right |
| 0 | 0 | *z* | 0* | 0.0001* | 0* | 0* | 0* | 0.0001* | 0.0002* | 0* | 0.0003* | 0* |
| | | *M* | 0.0007* | 0.0009* | 0.0012* | 0.0009* | 0.0013* | 0.0005* | 0.0013* | 0.0007* | 0.0012* | 0.0006* |
| 0 | .14 | *z* | 0.0003* | 0.0001* | 0.0013* | 0.0008* | 0.0023* | 0.0022* | 0.0064* | 0.0046* | 0.0129 | 0.0127 |
| | | *M* | 0.0024* | 0.0018* | 0.0063* | 0.0040* | 0.0079* | 0.0079* | 0.0079* | 0.0140 | 0.0194 | 0.0190 |
| 0 | .39 | *z* | 0.0047* | 0.0042* | 0.0116* | 0.0078* | 0.0201 | 0.0177 | 0.0218 | 0.0249 | 0.0256 | 0.0241 |
| | | *M* | 0.0129 | 0.0122* | 0.0153 | 0.0165 | 0.0214 | 0.0201 | 0.0228 | 0.0235 | 0.0249 | 0.0244 |
| 0 | .59 | *z* | 0.0126 | 0.0119* | 0.0189 | 0.0169 | 0.0211 | 0.0213 | 0.0266 | 0.0232 | 0.0248 | 0.0244 |
| | | *M* | 0.0202 | 0.0179 | 0.0205 | 0.0181 | 0.0211 | 0.0215 | 0.0263 | 0.0223 | 0.0248 | 0.0241 |
| Average | | *z* | 0.0044* | 0.0041* | 0.0080* | 0.0064* | 0.0109* | 0.0103* | 0.0138 | 0.0132 | 0.0159 | 0.0153 |
| | | *M* | 0.0091* | 0.0082* | 0.0108* | 0.0099* | 0.0129 | 0.0125 | 0.0146 | 0.0151 | 0.0176 | 0.0170 |

*Note.* *z* refers to confidence limits based on the traditional *z* test. *M* refers to the confidence limits formed based on the distribution of the product tables in Meeker et al. (1981). Values marked with * are outside Bradley (1978) robustness criteria.

Table 2
Proportion of True Values to Left and Right of 95% Confidence Intervals - No Zero Paths, Study 1

| Effect Size | | | Sample Size | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 50 | | 100 | | 200 | | 500 | | 1000 | |
| α | β | Test | true to left | true to right | true to left | true to right | true to left | true to right | true to left | true to right | true to left | true to right |
| .14 | .14 | *z* | 0.0006* | 0.0464* | 0.0019* | 0.0883* | 0.0040* | 0.0906* | 0.0070* | 0.0669* | 0.0112* | 0.0533* |
| | | *M* | 0.0064* | 0.0420* | 0.0098* | 0.0775* | 0.0133 | 0.0830* | 0.0177 | 0.0429* | 0.0189 | 0.0362 |
| .14 | .39 | *z* | 0.0037* | 0.0521* | 0.0062* | 0.0451* | 0.0098* | 0.0430* | 0.0138 | 0.0352 | 0.0191 | 0.0319 |
| | | *M* | 0.0134 | 0.0464* | 0.0177 | 0.0385* | 0.0198 | 0.0384* | 0.0208 | 0.0294 | 0.0204 | 0.0300 |
| .14 | .59 | *z* | 0.0072* | 0.0340 | 0.0115* | 0.0347 | 0.0172 | 0.0311 | 0.0189 | 0.0269 | 0.0219 | 0.0280 |
| | | *M* | 0.0165 | 0.0322 | 0.0230 | 0.0308 | 0.0258 | 0.0296 | 0.0195 | 0.0259 | 0.0219 | 0.0280 |
| .39 | .39 | *z* | 0.0062* | 0.0752* | 0.0109* | 0.0577* | 0.0121* | 0.0467* | 0.0158 | 0.0369 | 0.0165 | 0.0340 |
| | | *M* | 0.0165 | 0.0559* | 0.0192 | 0.0410* | 0.0190 | 0.0334 | 0.0197 | 0.0269 | 0.0166 | 0.0305 |
| .39 | .59 | *z* | 0.0089* | 0.0571* | 0.0104* | 0.0471* | 0.0146 | 0.0404* | 0.0186 | 0.0318 | 0.0195 | 0.0318 |
| | | *M* | 0.0193 | 0.0417* | 0.0183 | 0.0353 | 0.0222 | 0.0292 | 0.0187 | 0.0293 | 0.0195 | 0.0318 |
| .59 | .59 | *z* | 0.0094* | 0.0537* | 0.0137 | 0.0435* | 0.0152 | 0.0414* | 0.0168 | 0.0307 | 0.0211 | 0.0300 |
| | | *M* | 0.0197 | 0.0381* | 0.0219 | 0.0331 | 0.0185 | 0.0318 | 0.0168 | 0.0293 | 0.0211 | 0.0300 |
| Average | | *z* | 0.0060* | 0.0531* | 0.0091* | 0.0527* | 0.0122* | 0.0489* | 0.0152 | 0.0381* | 0.0182 | 0.0348 |
| | | *M* | 0.0153 | 0.0427* | 0.0183 | 0.0427* | 0.0198 | 0.0409* | 0.0189 | 0.0306 | 0.0197 | 0.0311 |

*Note.* *z* refers to confidence limits based on the traditional *z* test. *M* refers to the confidence limits formed based on the distribution of the product tables in Meeker et al. (1981). Values marked with * are outside Bradley (1978) robustness criteria.

The inaccuracy of some confidence limits based on the distribution of the product may also be due to the combination of the sampling variability of $\delta_\alpha$ and $\delta_\beta$ and the different shape of the distribution of the product for each different combination of $\delta_\alpha$ and $\delta_\beta$. For example, when the true population $\alpha$ is zero, the sample $\delta_\alpha$ is commonly nonzero owing to sampling variability. Because the shape of the distribution of the product changes depending on the value of $\delta_\alpha$, the confidence limits are subject to sampling variability. Nevertheless, even with sample values of $\delta_\alpha$ and $\delta_\beta$, the confidence limits based on the distribution of the product were more accurate than the method based on a normal distribution for the indirect effect.

Resampling approaches may yield more accurate confidence limits and also provide a test of the significance of the indirect effect (Manly, 1997; Noreen, 1989). The bootstrap resampling procedure is one method that may provide more accurate confidence limits (Bollen & Stine, 1990), but Type I error rates based on the standard percentile bootstrap approach are also lower than predicted for small values of $\delta_\alpha$ and $\delta_\beta$ (Lockwood & MacKinnon, 1998). The purpose of Study 2 is to evaluate several resampling methods that may yield more accurate confidence limits.

## *Study 2*

In Study 1, although confidence limits for the indirect effect were more accurate when the distribution of the product was taken into account, there were still cases where the number of times that the true value was outside the range of the confidence limits was smaller than expected. For example, the proportion outside the range for the case with $\alpha = 0$ and $\beta = 0$ was much lower than .025 for 95% confidence limits.

Several researchers have suggested that resampling methods such as the jackknife and the bootstrap may provide more accurate tests of the indirect effect (Bollen & Stine, 1990; Lockwood & MacKinnon, 1998; Shrout & Bolger, 2002). Bollen and Stine (1990) found that bootstrap confidence limits for the indirect effect were asymmetric. Lockwood and MacKinnon (1998) also obtained asymmetric confidence limits and presented a computer program to conduct the bootstrap confidence intervals for the indirect effect. Most recently, Shrout and Bolger (2002) recommended bootstrap methods to assess mediation for small to moderate sample sizes. Resampling methods are generally considered the method of choice when the assumptions of classical statistical methods are not met (Manly, 1997; Rodgers, 1999), such as for the nonnormal distribution of the indirect effect. Thus the purpose of Study 2 was to determine whether jackknife, Monte Carlo, and bootstrap resampling methods yield more

accurate confidence limits for the indirect effect than single sample methods used in Study 1.

## *Methods*

### *Simulation Description*

The simulation procedure in Study 1 was used in Study 2 with four exceptions: sample size, parameter combinations, number of replications, and resampling methods. First, only four sample sizes were simulated: 25, 50, 100, and 200. Because resampling methods are particularly useful when sample sizes are small, the two largest sample sizes from Study 1 were dropped and a sample size of 25 was added. Second, a subset of the combinations of parameter values were simulated to reduce the considerable computational demands of simulation studies of resampling methods. The ten combinations were $\alpha = 0$ $\beta = 0$, $\alpha = 0$ $\beta = .14$, $\alpha = 0$ $\beta = .39$, $\alpha = 0$ $\beta = .59$, $\alpha = \beta = .14$, $\alpha = \beta = .39$, $\alpha = \beta = .59$, $\alpha = .14$ $\beta = .39$, $\alpha = .14$ $\beta = .59$, and $\alpha = .39$ $\beta = .59$. These ten parameter combinations are the ones presented in the Tables for Study 1. Third, one thousand replications were conducted for each of the 40 combinations of sample size and parameters. Fourth, for each of the 40,000 (4 combinations of sample size times 10 parameter value combinations times 1000 replications) different data sets, six resampling methods were applied. For the bootstrap methods, a total of 1000 resampled data sets from each of the 40,000 data sets were used. That is, each bootstrap method entailed 1,000,000 (1000 replications times 1000 bootstrap samples) data sets for each of the 40 combinations of sample size and parameter values. For the jackknife method, the number of samples was the same as the sample size ($N$). Each of the resampling methods are described in more detail in the next section.

### *Confidence Limits*

Confidence limits for the methods described in Study 1 were formed in the same manner in Study 2. For the bootstrap methods, the confidence limits were obtained from the bootstrap distribution. For the jackknife method, a new jackknife estimate and standard error were used to compute confidence limits as described below. Confidence limits for the indirect effect were calculated for 80%, 90%, and 95% intervals. The proportion of times that the true value of the indirect effect was to the left of the lower confidence limit and to the right of the upper confidence limit was calculated for each method. Using the 95% confidence interval, the true indirect effect would

be predicted to be above the upper confidence limit 25 (.025 = 25/1000) times and below the lower confidence limit 25 (.025 = 25/1000) times. The additional confidence intervals were included to examine confidence limit performance at several levels and also because a greater number of true indirect effects will be outside the intervals for smaller intervals, for example, true indirect effects are expected outside the 90% confidence intervals 100 times. The liberal criterion described by Bradley (1978) was used for each of three different confidence limits, 95%, 90%, and 80% corresponding to intervals of .0125–.0375, .025–.075, and .05–.15, respectively. The total number of times the observed percentage was outside the robustness interval was tabulated for each combination of the three confidence intervals, four sample sizes, ten combinations of effect sizes, and upper and lower confidence intervals. A superior method would have fewer observed percentages outside the 240 robustness intervals formed by 3 confidence intervals times 10 parameter combinations times 4 sample sizes times 2 upper and lower confidence limits.

*Type I Error Rates and Statistical Power*

The observed Type I error rates and statistical power were also computed for each method. An effect was considered statistically significant if zero was not included in the confidence interval. For Type I error rates, the liberal Bradley (1978) robustness interval was also computed and Type I error rates outside the interval are indicated by an asterisk in the Tables.

*Single Sample Methods*

The $z$ test was calculated in the same way as in Study 1. The calculation of the $M$ test confidence limits was also the same as reported in Study 1 with one minor exception. The critical values for the $M$ test confidence limits in Study 2 come from an augmented table for the 95% confidence limits. The tabled values in Meeker et al. (1981) correspond to $\delta$ values in increments of .4 while the augmented table has values for $\delta$ values in increments of .2. These additional values were obtained with a FORTRAN algorithm written by Alan Miller which is a minor modification of the method in Meeker and Escobar (1994) and is available at http://users.bigpond.net.au/amiller (file name: fnprod.f90).

To address the possibility that the indirect effect is distributed in a manner different from the product of two standard normal variables a series of distributions were simulated to make a table of critical values similar to the

Meeker et al. (1981) table but with entries based on an empirical simulation rather than a theoretical distribution. This method is called the empirical-*M* method in this article. This new table of critical values may be more accurate if $\hat{\alpha}\hat{\beta}$ is distributed as *t* variates from regression analysis rather than *z* variates. Each combination of $\delta_\alpha$ and $\delta_\beta$ in Meeker et al. (1981) was used to generate 10,000 samples with 1000 observations each, from which the product term, $\hat{\delta}_\alpha\hat{\delta}_\beta$, was then computed for each sample. The distribution of $\hat{\delta}_\alpha\hat{\delta}_\beta$ was then used to find percentiles corresponding to upper and lower 95% confidence limits. The values were standardized so that they could be used for any sample size. The resulting table was then used in an identical manner to using the table in Meeker et al. (1981) with one exception. If sample values of $\delta_\alpha$ and $\delta_\beta$ were greater than 12, then the critical value for 12 was used from the empirical distribution.[2]

*Resampling Methods*

Six resampling methods were evaluated in this study: jackknife, percentile bootstrap, bias-corrected bootstrap, bootstrap-*t*, bootstrap-*Q*, and Monte Carlo. All of the methods adjust for nonnormal distributions although the bootstrap-*Q* and the bias-corrected bootstrap may be especially appropriate for severely nonnormal data (Chernick, 1999; Manly, 1997).

*Jackknife*. The jackknife (Mosteller & Tukey, 1977) is one of the first resampling methods described in the research literature. For a sample size *N*, there are *N* jackknife samples

$$s_{jackknife} = \sqrt{\frac{N-1}{N}\Sigma\left[\theta_{(i)} - \theta_{(.)}\right]^2} \tag{8}$$

formed by removing one observation at a time from the original sample, so that each jackknife sample has $N - 1$ observations. The jackknife estimate is the average estimate across the *N* jackknife samples. The standard error of the jackknife estimate is obtained using Equation 8, where $\theta_{(i)}$ is the value of the indirect effect in the $i^{th}$ jackknife sample and $\theta_{(.)}$ is the jackknife estimate of the statistic. Confidence limits were formed using Equation 5, substituting $s_{jackknife}$ for $\hat{\sigma}_{\hat{\alpha}\hat{\beta}}$ and $\theta_{(.)}$ for $\hat{\alpha}\hat{\beta}$.

*Percentile Bootstrap*. The basic bootstrap confidence limits were obtained with the percentile method as described by Efron and Tibshirani (1993). The

[2] The empirical-*M* critical values are given at our website given in Footnote 1.

sample parameter values at the ω/2 and 1 – ω/2 percentile of the bootstrap sampling distribution were used as the lower and upper confidence limits. For example, the percentile method 90% confidence limits would be the values of the bootstrap sampling distribution at 5% and 95% cumulative frequency.

*Bias-corrected Bootstrap*. The second bootstrap method corrects for bias in the central tendency of the estimate. This bias is expressed by $\hat{z}_0$, which is the $z$ score of the value obtained from the proportion of bootstrap samples below the original estimate in the total number of bootstrap samples taken. In other words $\hat{z}_0$ is the $z$ score of the percentile of the observed sample indirect effect. The upper confidence limit was then found as the $z$ score of $2\hat{z}_0 + z_{1-\omega/2}$ and the lower limit was $2\hat{z}_0 + z_{\omega/2}$.

*Bootstrap-t*. The bootstrap-$t$ method is based on the $t$ statistic rather than the indirect effect itself. It requires the standard error of the parameter estimate for each bootstrap sample which is the sampling standard deviation of the bootstrap sample. The value $T$ is formed for each bootstrap sample by dividing the difference between the bootstrap estimate and the original sample estimate by the bootstrap sample's standard error. The ω/2 and 1 – ω/2 percentiles of $T$ are then found. Confidence limits were formed as $(\hat{\alpha}\hat{\beta} - T_{1-\omega/2} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}}, \hat{\alpha}\hat{\beta} - T_{\omega/2} * \hat{\sigma}_{\hat{\alpha}\hat{\beta}})$

*Bootstrap-Q*. The bootstrap-$Q$ is a transformation of the bootstrap-$t$ that makes the distribution more closely follow the $t$ distribution (Manly, 1997). The bootstrap-$Q$ is obtained by transforming the bootstrap-$t$ using Equation 9 shown below where $s$ is skewness in each bootstrap distribution of $T$, $T$ is the bootstrap-$t$ value in each individual bootstrap sample, and $N$ is the sample size (Manly, 1997).

$$Q(T) = T + (sT^2)/3 + (s^2T^3)/27 + s/(6N) \tag{9}$$

Critical values of $Q$ are then found at the ω/2 and 1 – ω/2 percentiles. These critical values of $Q(T)$ are then transformed back to values of $T$ (using Equation 10 below which is Equation 3.14 in Manly, 1997) and these values of $W(Q) = T$ are used in an identical manner to the bootstrap-$t$ for the confidence limits.

$$W(Q) = 3[\{1 + s[Q - s/(6N)]\}^{1/3} - 1]/s \tag{10}$$

*Monte Carlo*. There are three major steps for the Monte Carlo method.

1. First, the indirect effect estimates, $\hat{\alpha}$ and $\hat{\beta}$, and standard errors, $\hat{\sigma}_{\hat{\alpha}}$ and $\hat{\sigma}_{\hat{\beta}}$ are estimated for the sample.

2. These sample estimates are then used to generate a sampling distribution of the product of $\alpha$ and $\beta$, based on generating a distribution of 1000 random samples with population values equal to the sample values, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_{\hat{\alpha}}$ and $\hat{\sigma}_{\hat{\beta}}$.

3. The lower and upper confidence limits for the indirect effect in each sample are the values in the generated distribution in Step 2, corresponding to the percentiles of the upper and lower confidence limits.

## *Results*

### *Confidence Limits*

Table 3 shows the proportion of times the true indirect effect was greater than or less than the confidence interval generated by each method for 95% confidence limits. In the interest of conserving space, we present the results only for the 95% confidence interval averaged across parameter values. The results for each parameter combination and for 90% and 80% confidence intervals are comparable and are available from the first author. The same pattern of results for the *M* and *z* confidence limits observed in Study 1 were obtained in Study 2. The *M* confidence limits are closer to the expected percentages than the confidence limits based on the normal distribution. For models where the true indirect effect equals 0, the percentages are almost always lower than .025, suggesting that the sample confidence intervals are too wide. Only the bias-corrected bootstrap has some percentages that are larger than the robustness interval, and this occurred most often for the $\alpha = 0$ $\beta = .39$ and $\alpha = 0$ $\beta = .59$ parameter combinations. Averaging across parameter value conditions for a true indirect effect equal to 0, only the bias-corrected bootstrap is never outside the robustness intervals. All other approaches tend to have percentages that are too small. Overall, percentages are more likely to be inside the robustness interval as effect size and sample size increase.

Table 4 summarizes the performance of the confidence limits by showing the number of times that the observed percentage was outside the robustness interval for each of the three confidence intervals and nine methods[3]. As shown in Table 4 for the 95% confidence interval, the observed percentage was outside the robustness interval 34-41 times for the two distribution of the

[3] Two resampling methods were investigated but are not reported in the text to conserve space and because the results were similar to methods in the article. The accelerated bias correction bootstrap method was also applied with the results comparable to but not better than the bias-corrected bootstrap. A method we labeled the *Q*-transform was also investigated in this research. The method consists of applying the transformation in Equation 9 but not transforming the *Q* back to a *T* using Equation 10. In general this method had somewhat better performance than the Bootstrap-*Q* although not as good as the bias corrected bootstrap.

product tests and all resampling methods except the jackknife. All of these methods were considerably better than the jackknife (59 times) which only had slightly better performance than the traditional $z$ test (61 times). Looking at the totals across all three confidence intervals, the bias-corrected bootstrap had the best performance with 73 times outside the robustness interval. The Empirical-$M$ (90), bootstrap percentile (97), bootstrap-$t$ (88), bootstrap-$Q$ (95) and Monte Carlo (92) all had generally similar performance. The $M$ test (114) appears to be intermediate between these more accurate methods and the traditional $z$ (144) and jackknife (138) which had the worst performance overall. Although not shown in the tables, the same pattern of results were also observed for squared bias (observed Type I error rate minus expected Type I error rate squared).

The Type I error rates and observed power of the single sample and resampling methods are shown in Table 5 for the .05 level of statistical significance. As in Table 3, the Type I error rates are always smaller than predicted for each method with the exception of the bias-corrected bootstrap which is also the only method that on average has rates that are not outside the robustness interval. With the exception of the traditional $z$ and the jackknife, all tests have Type I error rates that are not outside the robustness interval for sample size of at least 100. The bias-corrected method also had the most statistical power overall, followed by the bootstrap-$Q$, $M$, and Empirical-$M$, which had very similar overall power values. The percentile bootstrap, Monte Carlo, and bootstrap-$t$ test had very similar overall power. The traditional $z$ test and jackknife were similar and had considerably less power than the other methods.

*Discussion*

The different tests can be grouped in four general categories based on the results of Study 2. The confidence limits based on the jackknife and the traditional $z$ test are in the first group of tests which had the worst performance of all the tests with the least power and the lowest Type I error rates. The second group of tests consists of the percentile bootstrap, bootstrap-$t$ and Monte Carlo tests which have comparable power and confidence limits that tend to be too wide. The third group of tests, $M$ test, Empirical-$M$ test, and the bootstrap-$Q$, have more power and confidence limits that are not as wide as the tests in the second group. All of these tests had Type I errors that were never above the robustness interval, yet had more power than the methods in the first two groups. If a researcher wanted to avoid exceeding nominal Type I error rates, these are the methods of choice. The fourth group consists of the bias-corrected bootstrap which had

Table 3
Proportion of True Value to the Left and Right of 95% Confidence Intervals, Study 2

| | | Sample Size | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 25 | | 50 | | 100 | | 200 | |
| Indirect Effect | Method | left | right | left | right | left | right | left | right |
| Null Models | *z* | 0.0020* | 0.0028* | 0.0055* | 0.0043* | 0.0090* | 0.0083* | 0.0098* | 0.0078* |
| | *M* | 0.0103* | 0.0140 | 0.0113* | 0.0145 | 0.0180 | 0.0188 | 0.0183 | 0.0130 |
| | Empirical-*M* | 0.0098* | 0.0140 | 0.0128 | 0.0150 | 0.0188 | 0.0195 | 0.0188 | 0.0140 |
| | Jackknife | 0.0033* | 0.0033* | 0.0053* | 0.0063* | 0.0080* | 0.0083* | 0.0103* | 0.0090* |
| | Bootstrap percentile | 0.0090* | 0.0113* | 0.0140 | 0.0150 | 0.0188 | 0.0190 | 0.0195 | 0.0150 |
| | Bootstrap Bias-corrected | 0.0245 | 0.0268 | 0.0255 | 0.0260 | 0.0293 | 0.0330 | 0.0275 | 0.0275 |
| | Bootstrap-*t* | 0.0065* | 0.0088* | 0.0133 | 0.0105* | 0.0160 | 0.0180 | 0.0178 | 0.0138 |
| | Bootstrap-*Q* | 0.0075* | 0.0103* | 0.0125 | 0.0110* | 0.0165 | 0.0183 | 0.0175 | 0.0135 |
| | Monte Carlo | 0.0070* | 0.0108* | 0.0103* | 0.0113* | 0.0165 | 0.0153 | 0.0160 | 0.0110* |
| Non-zero Models | *z* | 0.0030* | 0.0547* | 0.0077* | 0.0577* | 0.0098* | 0.0598* | 0.0132 | 0.0480* |
| | *M* | 0.0120* | 0.0502* | 0.0200 | 0.0467* | 0.0192 | 0.0492* | 0.0198 | 0.0398* |
| | Empirical-*M* | 0.0118* | 0.0408* | 0.0192 | 0.0473* | 0.0190 | 0.0487* | 0.0190 | 0.0378* |
| | Jackknife | 0.0057* | 0.0528* | 0.0072* | 0.0570* | 0.0125 | 0.0582* | 0.0135 | 0.0487* |
| | Bootstrap percentile | 0.0127 | 0.0438* | 0.0187 | 0.0437* | 0.0233 | 0.0413* | 0.0222 | 0.0400* |
| | Bootstrap Bias-corrected | 0.0207 | 0.0553* | 0.0268 | 0.0498* | 0.0288 | 0.0430* | 0.0273 | 0.0340 |
| | Bootstrap-*t* | 0.0098* | 0.0352 | 0.0177 | 0.0372 | 0.0202 | 0.0357 | 0.0223 | 0.0350 |
| | Bootstrap-*Q* | 0.0185 | 0.0603* | 0.0273 | 0.0470* | 0.0297 | 0.0470* | 0.0265 | 0.0365 |
| | Monte Carlo | 0.0098* | 0.0295 | 0.0172 | 0.0317 | 0.0168 | 0.0350 | 0.0182 | 0.0335 |

*Note.* Values marked with * are outside Bradley (1978) robustness criteria.

Table 4
Number of Times UCL and LCL Proportions were Outside the Robustness Interval as a function of Method and Confidence Interval in Study 2

| | 80% | | | 90% | | | 95% | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Left | Right | Total | Left | Right | Total | Left | Right | Total | Left | Right | Total |
| *z* | 15 | 19 | 34 | 24 | 25 | 49 | 29 | 32 | 61 | 68 | 76 | 144 |
| *M* | 11 | 22 | 33 | 15 | 29 | 44 | 13 | 24 | 37 | 39 | 75 | 114 |
| Empirical-*M* | 9 | 14 | 23 | 12 | 19 | 31 | 12 | 24 | 36 | 33 | 57 | 90 |
| Jackknife | 14 | 16 | 30 | 23 | 26 | 49 | 28 | 31 | 59 | 65 | 73 | 138 |
| Bootstrap Percentile | 10 | 18 | 28 | 12 | 22 | 34 | 11 | 24 | 35 | 33 | 64 | 97 |
| Bootstrap Bias-corrected | 3 | 8 | 11 | 7 | 15 | 22 | 14 | 26 | 40 | 24 | 49 | 73 |
| Bootstrap-*t* | 8 | 13 | 21 | 12 | 17 | 29 | 16 | 22 | 38 | 36 | 52 | 88 |
| Bootstrap-*Q* | 8 | 13 | 21 | 12 | 21 | 33 | 15 | 26 | 41 | 35 | 60 | 95 |
| Monte Carlo | 11 | 13 | 24 | 15 | 19 | 34 | 16 | 18 | 34 | 42 | 50 | 92 |
| Total | 89 | 136 | 225 | 132 | 193 | 325 | 154 | 227 | 381 | 375 | 556 | 931 |

Table 5
Type I Error Rates and Power as a Function of Method and Sample Size, Study 2

| | | Sample Size | | | |
|---|---|---|---|---|---|
| Indirect Effect | Method | 25 | 50 | 100 | 200 |
| Null Models | *z* | 0.005* | 0.010* | 0.017* | 0.018* |
| | *M* | 0.024* | 0.026 | 0.037 | 0.031 |
| | Empirical-*M* | 0.024* | 0.028 | 0.038 | 0.033 |
| | Jackknife | 0.007* | 0.012* | 0.016* | 0.019* |
| | Bootstrap Percentile | 0.020* | 0.028 | 0.036 | 0.034 |
| | Bootstrap Bias -corrected | 0.051 | 0.052 | 0.064 | 0.055 |
| | Bootstrap-*t* | 0.015* | 0.024* | 0.034 | 0.032 |
| | Bootstrap-*Q* | 0.018* | 0.024* | 0.035 | 0.031 |
| | Monte Carlo | 0.018* | 0.024* | 0.030 | 0.029 |
| Non-zero Models | *z* | 0.119 | 0.339 | 0.544 | 0.674 |
| | *M* | 0.235 | 0.446 | 0.599 | 0.718 |
| | Empirical-*M* | 0.237 | 0.451 | 0.601 | 0.720 |
| | Jackknife | 0.120 | 0.326 | 0.535 | 0.672 |
| | Bootstrap Percentile | 0.195 | 0.418 | 0.584 | 0.708 |
| | Bootstrap Bias -corrected | 0.271 | 0.479 | 0.620 | 0.733 |
| | Bootstrap-*t* | 0.200 | 0.421 | 0.588 | 0.707 |
| | Bootstrap-*Q* | 0.233 | 0.448 | 0.608 | 0.722 |
| | Monte Carlo | 0.208 | 0.416 | 0.584 | 0.705 |

*Note.* Values marked with * are outside Bradley (1978) robustness criteria for Type I error rates.

slightly more power than methods in the second category and had the most accurate confidence intervals. The bias-corrected bootstrap did have Type I error rates that were above the robustness interval for some parameter combinations, but overall, the bias-corrected method had average Type I error rates within the robustness interval. As a result, the single best method overall was the bias-corrected bootstrap which had Type I error rates close to the nominal level along with more power than the other methods.

*Example*

The following example illustrates the methods used in this article with data from the Adolescents Training and Learning to Avoid Steroids (ATLAS) program. The ATLAS program is a multi-component program administered to high school football players to prevent use of anabolic androgenic steroids (AAS). More details of the program may be found in Goldberg et al. (1996) and single and multiple mediator model results using the standard $z$ significance test for the mediated effect may be found in MacKinnon et al. (2001). The data for this simplified example were from 861 cases (from 15 treatment schools and 16 control schools) with complete data on three variables, $X$-exposure to the program or not, $X_M$-perceived severity of anabolic steroid use, and $Y$-nutrition behaviors. One part of the program was designed to increase the perceived severity of using steroids, which was hypothesized to increase proper nutrition behaviors.

Confidence limits for the indirect effect were computed for three single sample tests: the traditional $z$, the $M$ test, and the empirical-$M$ test, as well as six resampling methods: the jackknife, percentile bootstrap, bootstrap-$t$, bootstrap-$Q$, bias-corrected bootstrap, and the Monte Carlo method. Ninety-five percent confidence limits were formed for the data using each method as described in Study 2. For the jackknife, 861 samples were drawn from the original data, each one excluding a different case. All bootstrap methods were performed with 1000 resamples from the original data. Additionally, 1000 Monte Carlo samples were generated using the observed estimates of $\alpha$, $\beta$, $\sigma_\alpha$, and $\sigma_\beta$ (.2731, .0736, .0894, and .0300 respectively).

The indirect effect equaled .0201 with a standard error of .0105 using Equation 4. These values were used with Equation 5 to find upper and lower confidence limits for the $z$ method. The $M$ test limits were found using the sample values of $\delta_\alpha = 3.0539$ and $\delta_\beta = 2.4486$. Critical values were then found in the augmented Meeker et al. (1981) tables using rounded values of 3 and 2.4. The upper critical $M$ value was 2.2683 and the lower was equal to –1.5969. These values were used in Equations 6 and 7 to find the upper and lower $M$ test confidence limits. Table 6 presents the upper and lower confidence limits of the mediated effect using these single sample methods and the resampling tests. The results demonstrate many of the findings of the simulation studies. First, zero is included in the confidence limits for the traditional $z$ method, which would lead to the conclusion that the indirect effect is nonsignificant. When the more powerful $M$ test is used however, the effect is statistically significant. The resampling approaches also suggest a significant indirect effect.

Table 6
Upper and Lower Indirect Effect Confidence Limits for ATLAS Data

| Method | Lower CL | Upper CL |
|---|---|---|
| *z* | –0.0005 | 0.0407 |
| *M* | 0.0033 | 0.0407 |
| Empirical-*M* | 0.0025 | 0.0434 |
| Jackknife | 0.0002 | 0.0400 |
| Percentile bootstrap | 0.0020 | 0.0380 |
| Bias-corrected bootstrap | 0.0073 | 0.0532 |
| Bootstrap-*t* | 0.0003 | 0.0400 |
| Bootstrap-*Q* | 0.0003 | 0.0400 |
| Monte Carlo | 0.0017 | 0.0441 |

## *General Discussion*

The purpose of this article was to evaluate two alternatives to improve confidence limit coverage for the indirect effect. One method incorporated the distribution of the product to construct confidence limits and the second method used resampling approaches that make fewer assumptions about the distribution of the indirect effect. Both methods led to improved confidence limit coverage compared to the traditional method based on the normal distribution. In Study 1, the confidence limit coverage for the method based on the distribution of the product was nearly always better than the traditional method for a large number of combinations of effect size. In Study 2, resampling methods had better performance than the method based on the normal distribution, with the exception of the jackknife. Not all resampling methods were superior to the distribution of the product methods, however, which suggests caution in selecting a resampling method. The best single sample tests are based on the confidence limits using *M* from the Meeker et al. (1981) tables or the empirical-*M* confidence limits based on empirically generated critical values. Both use the distribution of the product to create confidence limits and test the significance of the results. Either of these tests are the single sample method of choice. The empirical-*M* test did have slightly better performance than the *M* test in terms of times outside the robustness interval but overall power and Type I error rates were very similar. If a researcher does not have the raw data available for analysis,

resampling methods described in this article cannot be conducted and these single sample methods are the only methods available.

The bias-corrected bootstrap is the method of choice if resampling methods are feasible. There are limitations to the use of resampling methods, however. These include the lack of statistical software to conduct the analysis and the increased computational time required for the analysis. The bias-corrected bootstrap is included as an option in the AMOS (Arbuckle & Wothke, 1999) covariance structure analysis program. The EQS program and LISREL programs can be used to conduct resampling analyses but the user must write a separate program to compute the bootstrap method after these programs write out the results of each individual bootstrap sample. All of the resampling tests used in this article are now included in an updated version of the SAS program in Lockwood and MacKinnon (1998). Computation time does not seem like a reasonable argument for not using resampling methods because of the speed of current computing, although relatively large models may take considerable time.

Another potential limitation of resampling methods is that the difference between resampling methods and single sample methods is small in many cases. In fact, Bollen and Stine (1990) report percentile bootstrap confidence limits and not bias-corrected bootstrap confidence limits because their limits were so similar. In the present study for example, the bias corrected bootstrap and the $z$ test came to the different conclusions regarding whether the population value was inside or outside the confidence interval only 5.5% of the time across all confidence intervals examined in this study. The discrepancy between the bias-corrected bootstrap and the $M$ test was only 4.5%. As shown in the example data, the confidence limits were very similar across the methods, although only the traditional $z$ included zero in the confidence interval. There is some evidence from this study that for cases where the values of either $\delta_\alpha$ and $\delta_\beta$ are large, resampling methods do not provide much improvement over single sample methods. The additional effort required for resampling methods may not be justified in these cases. When confidence limits are required for rather small values of $\delta_\alpha$ and $\delta_\beta$, then the resampling methods are more accurate than single sample tests. A final limitation of resampling tests has been called the first law of statistical analysis (Gleser, 1996). The law requires that any two statisticians analyzing the same data set with the same methods should come to identical conclusions. With resampling methods (other than the jackknife), it is possible that different conclusions could be reached because the resampled data sets generated by different statisticians will differ.

The results of this study highlight a difference between testing significance based on critical values versus confidence limits. There is a problematic result

of testing the significance of the indirect effect with the distribution of the product using the critical value for the distribution for $\alpha = 0$ and $\beta = 0$, which is 2.18 for a .05 Type I error rate. In this situation, either $\alpha$ or $\beta$ can be nonsignificant but the test based on the distribution of the product may be significant, indicating that the indirect effect is larger than expected by chance alone while one of the regression coefficients contributing to its effect is not. The statistical test of whether $\alpha = 0$ and $\beta = 0$ is more likely to be judged significant when the true values are $\alpha = 0$ and $\beta \neq 0$ or $\alpha \neq 0$ and $\beta = 0$, because the distribution of the indirect effect with these parameter values differs from the distribution for $\alpha = \beta = 0$. By selecting the critical value of 2.18 from the distribution of the product for $\alpha = 0$ and $\beta = 0$, the null hypothesis is $H_0$: $\alpha = 0$ and $\beta = 0$, which is rejected in three situations: when $\alpha \neq 0$ and $\beta \neq 0$, $\alpha = 0$ and $\beta \neq 0$, or $\alpha \neq 0$ and $\beta = 0$. The traditional $z$ test is a test of the null hypothesis $H_0$: $\alpha\beta = 0$, which is rejected only when $\alpha \neq 0$ and $\beta \neq 0$ but the assumption of the $z$ distribution does not appear to be accurate for the product of $\alpha$ and $\beta$, altering the Type I error rates and statistical power of this test. The *M* test of significance based on confidence limits also tests whether $\alpha\beta = 0$ and requires different critical values for the upper and lower limits. In this case a test of $H_0$: $\alpha = 0$ and $\beta = 0$ yields a different result than the test based on confidence limits.

There are other methods to evaluate indirect effects. These include the steps mentioned in Baron and Kenny (1986) and Judd and Kenny (1981) and the joint significance test of $\alpha$ and $\beta$ described in MacKinnon et al. (2002), which do not include explicit methods to compute confidence limits. There are other methods to compute confidence limits for the indirect effect based on the standard error of the difference in coefficients, $\tau - \tau'$ (e.g., Allison, 1995b ; Clogg, Petkova, & Cheng, 1995; Clogg, Petkova, & Shihadeh, 1992; Olkin & Finn, 1995) but these methods perform similar to the traditional $z$ test described in this article, in part because a normal distribution for the indirect effect is assumed.

There are several limitations of this article. The results may not extend beyond the single indirect effect model investigated. The influence of different distributions of $X$, $X_M$, and $Y_O$ was not evaluated. However, Study 1 included a binary independent variable as another condition. Because the results were virtually identical to the results for a continuous independent variable, they were not reported in this article. Use of the *M* test confidence limits for indirect effects consisting of the product of three or more paths would require the use of the distribution of the product of three or more random variables. There are analytical solutions for this distribution (Springer, 1979) but tabled values of the distribution are not available. The resampling methods can be applied to these more complicated models.

This article has also been silent regarding important conceptual issues in interpreting indirect effects. Here it is assumed that the indirect effect model is known and that $X$ precedes $X_M$ and $X_M$ precedes $Y_O$. In practice, the hypothesized chain of effects in an indirect effect may be wrong and there may be several equivalent models that will explain the relations equally well. For example, the mediating variable may actually change the independent variable that may then affect the dependent variable. In the case of a randomized experiment, the independent variable improves interpretation because it must precede the mediating variable and the dependent variable, but even in this situation the interpretation of indirect effects is more complicated than what might be expected (Holland, 1988; Sobel, 1998). Issues regarding the specificity of the effect to one or a few of many mediating variables and future experiments targeted at specific mediating variables improve the veracity of indirect effects (West & Aiken, 1997). None of these methods to test the statistical significance or compute confidence limits for indirect effects answer these critical conceptual questions, but when combined with careful replication studies these relations are clarified.

There are several ways that the confidence limits described in this article may be further improved. One approach is to use more extensive resampling methods such as bootstrapping residuals, iterated bootstrap, or methods to compute confidence limits based on the permutation test (Manly, 1997). More extensive tables for the *M* test critical values may also improve the performance of this method. Analytical work on the distribution of the product of two regression coefficients especially for small sample sizes may also lead to more accurate confidence limits.

The practical implication of the results of this article is that the traditional *z* test confidence limits can be substantially improved by using a method such as the *M* test that incorporates the distribution of the product of two normal random variables. The bias-corrected bootstrap provided the most accurate confidence limits and greatest statistical power, and is the method of choice if it is feasible to conduct resampling methods.

## References

Ajzen, I. & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice Hall.

Allison, P. D. (1995a). Exact variance of indirect effects in recursive linear models. *Sociological Methodology, 25*, 253-266.

Allison, P. D. (1995b). The impact of random predictors on comparisons of coefficients between models: Comment on Clogg, Petkova, and Haritou. *American Journal of Sociology, 100*, 1294-1305.

Alwin, D. F. & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*, 37-47.

Arbuckle, J. L. & Wothke, W. (1999). *Amos 4.0 users' guide version 3.6*. Chicago: Smallwaters.

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics,18*, 265-271.

Aroian, L. A., Taneja, V. S., & Cornwell, L. W. (1978). Mathematical forms of the distribution of the product of two normal variables. *Communications in Statistics: Theory and Methods, A7*, 165-172.

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

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.

Bentler, P. M. (1997). *EQS for Windows (Version 5.6)* [Computer program]. Encino, CA: Multivariate Software, Inc.

Bollen, K. A. & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115-140.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley.

Clogg, C. C., Petkova, E., & Cheng, T. (1995). Reply to Allison: More on comparing regression coefficients. *American Journal of Sociology, 100*, 1305-1312.

Clogg, C. C., Petkova, E., & Shihadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *Journal of Educational Statistics, 17*, 51-74.

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

Craig, C. C. (1936). On the frequency function of *xy*. *Annals of Mathematical Statistics, 7*, 1-15.

Duncan, O. D., Featherman, D. L., & Duncan, B. (1972). *Socioeconomic background and achievement*. New York: Seminar Press.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Gleser, L. J. (1996). Comment on "Bootstrap Confidence Intervals" by T. J. DiCiccio and B. Efron. *Statistical Science, 11,* 219-221.

Goldberg, L., Elliot, D., Clarke, G. N., MacKinnon, D. P., Moe, E., Zoref, L., et al. (1996). Effects of a multi-dimensional anabolic steroid prevention intervention: The Adolescents Training and Learning to Avoid Steroids (ATLAS) program. *Journal of the American Medical Association, 276*, 1555-1562.

Hansen, W. B. & Graham, J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine, 20,* 414-430.

Hanushek, E. A. & Jackson, J. E. (1977). *Statistical methods for social scientists*. New York: Academic Press.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hawkins, J. D., Catalano, R. F. & Miller, J. Y. (1992). Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychological Bulletin, 112*, 64-105.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology, 18,* 449-484.

Hoyle, R. H. & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195-222). Thousand Oaks, CA: Sage.

Hyman, H. H. (1955). *Survey design and analysis: Principles, cases and procedures*. Glencoe: The Free Press.

James, L. R. & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology, 69,* 307-321.

Jöreskog, K. G. & Sörbom, D. (1993). *LISREL (Version 8.12)* [Computer program]. Chicago: Scientific Software International.

Judd, C. M. & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602-619.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94,* 1372-1381.

Lockwood, C. M. & MacKinnon, D. P. (1998). Bootstrapping the standard error of the mediated effect. In *Proceedings of the Twenty-third Annual SAS Users Group International Conference* (pp. 997-1002). Cary, NC: SAS Institute.

Lomnicki, Z. A. (1967). On the distribution of products of random variables. *Journal of the Royal Statistical Society, 29*, 513-524.

MacKinnon, D. P. (1994). Analysis of mediating variables in prevention and intervention research. *National Institute on Drug Abuse Research Monograph Series*, *139*, 127-153.

MacKinnon, D. P. & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 17*, 144-158.

MacKinnon, D. P., Goldberg, L., Clarke, G. N., Elliot, D. L., Cheong, J., Lapin, A., et al. (2001). Mediating mechanisms in a program to reduce intentions to use anabolic steroids and improve exercise self-efficacy and dietary behavior. *Prevention Science, 2*, 15-28.

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R. & Wang, E. Y. I. (1991). Mediating mechanisms in a school-based drug prevention program: First year effects of the Midwestern Prevention Project. *Health Psychology, 10,* 164-172.

MacKinnon, D. P., Lockwood C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test the significance of mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30,* 41-62.

Manly, B. F. (1997). *Randomization and Monte Carlo methods in biology*. New York: Chapman and Hall.

McDonald, R. P. (1997). Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research, 32,* 1-38.

Meeker, W. Q., Jr., Cornwell, L. W., & Aroian, L. A. (1981). *Selected tables in mathematical statistics, volume VII: The product of two normally distributed random variables*. Providence, RI: American Mathematical Society.

Meeker, W. Q., Jr., & Escobar, L. A. (1994). An algorithm to compute the cdf of the product of two normal random variables. *Communications in Statistics: Simulation and Computation*, *23*, 271-280.

Mosteller, F. & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses: An introduction*. New York: John Wiley & Sons.

Olkin, I. & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118*, 155-164.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34*, 441-456.

SAS Institute. (1989). *SAS (Version 6.12)* [Computer program]. Cary, NC: Author.

Schoenberg, R. & Arminger, G. (1996). *LINCS (Version 2.0)* [Computer program]. Maple Valley, WA: Aptech Systems, Inc.

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

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290-312.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. *Sociological Methodology, 16*, 159-186.

Sobel, M. E. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods & Research, 16*, 155-176.

Sobel, M. E. (1990). Effect analysis and causation in linear structural equation models. *Psychometrika, 55*, 495-515.

Sobel, M. E. (1998). Causal inference in statistical models of the process of socioeconomic achievement: A case study. *Sociological Methods and Research, 27*, 318-348.

Springer, M. D. (1979). *The algebra of random variables*. New York: John Wiley and Sons.

Springer, M. D. & Thompson, W. E. (1966). The distribution of independent random variables. *SIAM Journal on Applied Mathematics, 14,* 511-526.

Stone, C. A. & Sobel, M. E. (1990). The robustness of estimates of total indirect effects in covariance structure models estimated by maximum likelihood. *Psychometrika, 55,* 337-352.

Tobler, N. S. (1986). Meta-analysis of 143 adolescent drug prevention programs: Quantitative outcome results of program participants compared to a control or comparison group. *Journal of Drug Issues, Fall*, 537-567.

Turner, R. J., Wheaton, B., & Lloyd, D. A. (1995). The epidemiology of social stress. *American Sociological Review, 60*, 104-125.

Warner, B. D. & Rountree, P. W. (1997). Local social ties in a community and crime model: Questioning the systemic nature of informal social control. *Social Problems, 44*, 520-536.

West, S. G. & Aiken, L. S. (1997). Towards understanding individual effects in multiple component prevention programs: Design and analysis strategies. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 167-209). Washington, DC: American Psychological Association.

Wolchik, S. A., Ruehlman, L. S., Braver, S. L., & Sandler, I. N. (1989). Social support of children of divorce: Direct and stress buffering effects. *American Journal of Community Psychology, 17,* 485-501.

*Accepted June, 2003.*