## Contents

**A Generalized Solution for Approximating the Power to Detect Effects of Categorical Moderator Variables Using Multiple Regression**  
Herman Aguinis  
Robert J. Boik  
Charles A. Pierce  
291

**Testing Interaction Effects in LISREL: Examination and Illustration of Available Procedures**  
Jose M. Cortina  
Gilad Chen  
William P. Dunlap  
324

**Single-Item Reliability: A Replication and Extension**  
John P. Wanous  
Michael J. Hudy  
361

**A Community of Voices: Using Allegory as an Interpretive Device in Action Research on Organizational Change**  
Joseph W. Grubbs  
376

**Index**  
393
A Generalized Solution for Approximating the Power to Detect Effects of Categorical Moderator Variables Using Multiple Regression

HERMAN AGUINIS  
University of Colorado at Denver

ROBERT J. BOIK  
CHARLES A. PIERCE  
Montana State University

Investigators in numerous organization studies disciplines are concerned about the low statistical power of moderated multiple regression (MMR) to detect effects of categorical moderator variables. The authors provide a theoretical approximation to the power of MMR. The theoretical result confirms, synthesizes, and extends previous Monte Carlo research on factors that affect the power of MMR tests of categorical moderator variables and the low power of MMR in typical research situations. The authors develop and describe a computer program, which is available on the Internet, that allows researchers to approximate the power of MMR to detect the effects of categorical moderator variables given user-input information (e.g., sample size, reliability of measurement). The approximation also allows investigators to determine the effects of violating certain assumptions required for MMR. Given the typically low power of MMR, researchers are encouraged to use the computer program to approximate power while planning their research design and methodology.

Researchers in numerous organization studies disciplines are interested in estimating interactive effects involving a categorical and a continuous variable. For example, differential prediction is operationalized as an interaction between ethnicity (e.g., minor-

Authors’ Note: This research was supported in part by a grant from the Graduate School of Business Administration (University of Colorado at Denver) to Herman Aguinis. A previous version of this article was presented as part of a symposium (J. M. Cortina, chair) at the meeting of the Society for Industrial and Organizational Psychology, Dallas, TX, April 1998. Portions of the research reported herein were conducted while Herman Aguinis was on sabbatical leave from the University of Colorado at Denver and holding visiting appointments at China Agricultural University–The International College of Beijing (People’s Republic of China), City University of Hong Kong (People’s Republic of China), Nanyang Technological University (Singapore), University of Science Malaysia (Penang, Malaysia), and University of Santiago de Compostela (Spain). Correspondence and reprint requests should be addressed to Herman Aguinis, Graduate School of Business Administration, University of Colorado at Denver, Campus Box 165, P.O. Box 173364, Denver, CO 80217-3364.

Organizational Research Methods, Vol. 4 No. 4, October 2001 291-323  
© 2001 Sage Publications
ity, nonminority) and test scores (e.g., general cognitive ability) on a measure of performance (e.g., supervisory ratings) (Cleary, 1968). If differential prediction is found, ethnicity is labeled a “moderator” of the relationship between test scores and performance. Gender is another illustration of a categorical variable whose moderating effect is of interest in several organization studies disciplines. Numerous additional examples were reviewed recently by Aguinis, Beaty, Boik, and Pierce (2000).

**Moderated Multiple Regression (MMR)**

The moderating effect of a categorical variable on the relationship between a continuous predictor and a continuous criterion is typically estimated using MMR. Assume that $X$ is a continuous predictor variable (e.g., preemployment test scores), $Y$ a continuous criterion variable (e.g., supervisory ratings of performance), and $Z$ a categorical predictor variable hypothesized to be a moderator (e.g., gender, dummy coded $1 =$ men and $2 =$ women). Equation 1 shows the linear regression model for predicting $Y$ from $X$, $Z$, and the interaction between $X$ and $Z$ (i.e., moderating effect of $Z$) represented by the $X \cdot Z$ product term (Aiken & West, 1991; Cohen & Cohen, 1983; Saunders, 1956):

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + \varepsilon,$$

where $\beta_0$ is the intercept, $\beta_1$ is the regression coefficient for $X$, $\beta_2$ is the regression coefficient for $Z$, $\beta_3$ is the regression coefficient for the product term that carries information about the interaction between $X$ and $Z$, and $\varepsilon$ is a normally distributed random error term. Rejecting the null hypothesis $H_0$ that the product term’s regression coefficient is zero indicates the presence of a moderating or interaction effect. Stated differently, rejecting this null hypothesis indicates that the regression of $Y$ on $X$ is unequal across levels of $Z$ (e.g., male and female subgroups). Note that although this illustration addresses a binary moderator variable (i.e., two levels), the MMR model allows the categorical moderator to take on any number of levels (e.g., a moderator with three levels could be ethnicity coded with African American, Latino/Latina, and White categories).

**Statistical Power Problems With MMR**

MMR is widely used to test hypotheses regarding the effects of categorical moderator variables in organization studies (e.g., Bobko & Russell, 1994), as well other fields including education (Aguinis, Nesler, Quigley, Lee, & Tedeschi, 1996), marketing (e.g., Mason & Perreault, 1991), and sociology (e.g., Smith, & Sasaki, 1979), among others.

Despite its pervasive use, researchers have lamented the low statistical power of MMR for nearly three decades (Aguinis, 1995; Aguinis, Bommer, & Pierce, 1996; McClelland & Judd, 1993; Zedeck, 1971). When MMR analyses are conducted at low levels of statistical power, researchers who fail to find support for their hypotheses regarding moderating effects do not know whether (a) their hypotheses are incorrect, or (b) their hypotheses are correct, but they failed to detect the moderating effect. This situation creates great uncertainty in theory development involving hypothesized moderating effects. Uncertainty regarding the presence of moderating effects is particularly serious because moderator variables are considered to be “at the very heart of
the scientific enterprise” (Hall & Rosenthal, 1991, p. 447) and to serve as indicators of a field’s scientific advancement and maturity (Hall & Rosenthal, 1991).1

The following factors have been identified as culprits for the typical low power of MMR: (a) reduced variance in the predictor variables (McClelland & Judd, 1993, also referred to as “range restriction” by Aguinis & Stone-Romero, 1997), (b) error variance heterogeneity across moderator-based subgroups (Alexander & DeShon, 1994), (c) measurement error (Busemeyer & Jones, 1983), (d) small total sample size (e.g., Alexander & DeShon, 1994), and (e) unequal sample size across the moderator-based subgroups (Stone-Romero, Alliger, & Aguinis, 1994) (see Aguinis, 1995, and Aguinis & Pierce, 1998a, for reviews). In practice, however, researchers often have little control over the size of their samples or the reliability levels of their measures. Thus, it is often the case that researchers conduct their MMR analyses under low power conditions. As a result, researchers may erroneously conclude that there is no moderating effect.

Given that most of the factors known to affect the power of MMR are not under the control of researchers and that researchers may not be aware of the low power of MMR in the sample at hand, Aguinis and colleagues (Aguinis & Pierce, 1998b; Aguinis, Pierce, & Stone-Romero, 1994) developed computer programs to estimate the power of MMR. These programs allow researchers to estimate the power of MMR tests for specific situations (e.g., large vs. small moderating effect magnitude). Despite the fact that these programs are available and useful to researchers in the quest for moderating effects, they suffer from five limitations. These limitations exist because the programs are based on algorithms derived from empirical (i.e., Monte Carlo) studies (Aguinis & Pierce, 1998b, is based on Aguinis & Stone-Romero, 1997; and Aguinis et al., 1994, is based on Stone-Romero et al., 1994).

The first limitation is that these programs do not include all the factors known to affect the power of MMR. More specifically, they assume that the measures of Y, X, and Z are free from measurement error. This is a tenable assumption for the categorical moderator Z but untenable for the predictor X and the criterion Y.

The second limitation of these programs is that the Monte Carlo studies on which they are based included only a limited range of values for factors affecting the power of MMR. For instance, Aguinis and Stone-Romero (1997) used values for total sample sizes of 60 and 300. Thus, the program by Aguinis and Pierce (1998b) based on the simulation results by Aguinis and Stone-Romero might not provide accurate power estimates for situations having sample size values far from 60 or 300. In addition, the Aguinis and Stone-Romero simulation did not include a negative relationship between X and Y for any of the moderator-based subgroups. Although these relationships are typically positive in the context of educational and preemployment testing, this is not the case in other research areas. For example, a health psychologist might want to test the moderating effect of gender on the negative relationship between the predictor “optimism” and the criterion “time to recover from surgery.” As will be illustrated later in this article, using computer programs based on Monte Carlo simulation results in which certain value ranges were not included (e.g., negative correlations) can lead to inaccurate power estimates.

The third limitation of the currently available programs to compute the power of MMR is that restriction on X is assumed to take on only the simplest form of truncation. That is, the programs allow users to specify whether X scores are truncated at a
specific cutoff point. This truncation is known to lower the power of MMR when \( X \) is normally distributed (Aguinis & Stone-Romero, 1997). However, simple truncation is only one of many situations that result in sample variances being smaller than population variances. For instance, in many situations, the probability that an individual will be selected in a sample does not depend only on his or her standing on variable \( X \) but also on his or her standing regarding other measured and unmeasured variables (Aguinis & Whitehead, 1997). More precisely, in educational and preemployment testing situations, many of the top scorers turn down an offer, so there might not be a precise score on \( X \) above which individuals have a higher probability of being included in the sample (Murphy, 1986). In short, differences between the sample and population variances for \( X \) might be due to a more complex sampling restriction mechanism, and the available programs to compute power do not allow for this perhaps more frequent research scenario.

The fourth limitation is that investigators can only compute power in situations in which the categorical moderator has two levels. This is a useful feature for researchers interested in testing the effects of binary moderator variables such as gender. However, there are many research situations in which the categorical moderator takes on more than two values (e.g., ethnicity coded using African American, Latino/Latina, and White categories). Although two levels are very frequently observed, researchers might also be interested in investigating the effects of moderators with more than two categories. However, at present there is no tool available to compute the power of MMR in these research situations.

Finally, the fifth limitation of existing programs is that they include situations with one continuous predictor \( X \) and one categorical moderator. Thus, existing programs do not allow researchers to compute the power of MMR in situations including two or more categorical moderator variables. One such common situation is an MMR model including the categorical moderators ethnicity and gender (cf. Aguinis et al., 2000). In such situations, a researcher might wish to compute the power to detect the \( X \) by ethnicity interaction, the \( X \) by gender interaction, and the ethnicity by gender interaction. In addition, if an interaction is expected, then a researcher might wish to compute the power to detect interaction contrasts among the \( Y \) on \( X \) slopes (see Boik, 1979, 1993, for a discussion of main and interaction effect contrasts). Computing power for these types of effects is not possible with the programs available at present.

**Present Study**

Given the limitations of previous efforts to produce highly accurate estimates of the power of MMR to detect the effects of categorical moderator variables, the goal of the present article is to develop a theory-based solution for approximating power. Overcoming limitations of previous empirically based research, we offer a theoretical result that allows researchers to approximate the power of MMR. First, we describe the theoretical approximation. Second, we present results of a Monte Carlo simulation showing its accuracy. Third, we describe a user-friendly computer program (MMRPOWER) available on the World Wide Web that we developed to implement the theoretical power approximation given user-input values. Thus, MMRPOWER allows users to compute power given their precise expected or actual situation (e.g., total sample size, sample sizes in each of the moderator-based subgroups, measurement error
for X and Y in each of the moderator-based subgroups). Finally, we use MMRPOWER to generate values for a number of typical situations in organization studies to examine the relative effects of the various factors that influence power.

In the sections that follow, we use the term **power** as a synonym of **power function**, which is the probability of rejecting the null hypothesis $H_0$ given specified values for all of the parameters (Casella & Berger, 2002). When $H_0$ is true and the homogeneity of error variance assumption is satisfied, then power equals the nominal preset rejection rate $\alpha$ (also referred to as Type I error rate or test size). However, when $H_0$ is true and the homogeneity assumption is not satisfied, then power can be greater or smaller than nominal $\alpha$. Thus, as we discuss later in more detail, MMR should not be used when the homogeneity of error variance assumption is violated (Aguinis, Petersen, & Pierce, 1999; Aguinis & Pierce, 1998a).

### An Analytic Approximation to the Power of MMR

#### Assumptions of the Model

Appendix A describes the assumptions used in developing the analytic approximation. Briefly, it is assumed that either X and/or Y could be measured with error. Regardless of the reliability of X and Y, it is assumed that a normal linear regression model holds for Y conditional on X, where Y and X are the observable scores rather than the true scores. This differs from the conventional errors-in-variables model in which $Y_{true}$ conditional on $X_{true}$ is assumed to follow a regression model (Brown & Fuller, 1990; Carroll & Ruppert, 1995). The MMR model is more appropriate when interest is in the relationship among the observable X and Y scores.

The analytic approximation allows for sampling restrictions that lead to a difference between the expected sample variance of X and the population variance of X. We use the term **variance multiplying factor** to refer to the expected sample variance of X divided by population variance of X. This ratio need not be the same in each moderator-based subgroup. Variance multiplying factors that differ from 1 can arise because of truncation (i.e., scores are included in the sample only if they are above or below a specific cutoff point) or other sampling restrictions. If X is normally distributed, then MMRPOWER will compute the variance multiplying factor. If X is not normally distributed, or if a sampling restriction other than truncation holds, then the user needs to input a value for the variance multiplying factor.

#### Factors Affecting Power

Appendix B provides a technical presentation of the null hypothesis, the MMR model, and the F statistic used in assessing the presence of a moderating effect in MMR. Appendix B defines the various components that are used in deriving the analytic approximation to power in Appendix C. Appendix C presents (a) the distribution of the F statistic conditional on X and (b) an approximation to the unconditional distribution of the F statistic. Note that Gatsionis and Sampson (1989) discussed the distinction between conditional and unconditional power in a simpler model than that considered here.
Theorem 2 in Appendix C gives an analytic expression for the power of the MMR $F$ test. The expression is technical, but this does not limit the practical usefulness of the theorem. The required computations are readily performed using MMRPOWER (see the Computer Program section below).

An examination of the analytic power approximation given in Theorem 2 reveals that the power of the MMR $F$ test depends on the following quantities: (a) preset nominal test size, $\alpha$; (b) number of moderator-based subpopulations, $k$; (c) sample sizes across moderator-based subgroups, $n_j$ for $j = 1, \ldots, k$; (d) difference in slopes of $Y$ on $X$ across moderator-based subpopulations, $\beta_j - \beta_k$ for $j = 1, \ldots, k - 1$; (e) reliabilities (e.g., Cronbach’s alpha, split-half, test-retest) for $Y$ in the $k$ moderator-based subpopulations, $\alpha_{y,j}$ for $j = 1, \ldots, k$; (f) reliabilities for $X$ in the $k$ moderator-based subpopulations, $\alpha_{x,j}$ for $j = 1, \ldots, k$; (g) correlations between $X$ and $Y$ in each of the moderator-based subpopulations, $\rho_j$ for $j = 1, \ldots, k$; (h) marginal variances of $Y$ in the $k$ subpopulations, $\sigma^2_{y,j}$, for $j = 1, \ldots, k$; (i) variance of $X$ in the $k$ subpopulations, $\sigma^2_{x,j}$ for $j = 1, \ldots, k$; and (j) ratio of expected sample variance of $X$ to population variance of $X$. Note, however, that some of these factors are not independent. For example, slopes are functions of reliabilities, correlations, marginal variances of $Y$, and variances of $X$.

That is, for slopes to change, at least one of these factors also needs to change.

Theorem 2 provides a theory-based synthesis of previous research regarding the variables that affect the power of MMR. For instance, consider the following variables: (a) sampling restriction on $X$ (Aguinis & Stone-Romero, 1997), (b) measurement error (e.g., Busemeyer & Jones, 1983), and (c) unequal sample sizes across moderator-based subgroups (Stone-Romero et al., 1994). Sampling restriction is incorporated by allowing the expected sample variance of $X$ to differ from the population variance of $X$. Stated differently, the analytic solution goes beyond simple truncation (e.g., an individual is included in the sample if his or her $X$ score is above a specific cutoff point) and addresses the more general issue that differences between sample and population variances affect power (McClelland & Judd, 1993). Measurement error is incorporated by allowing reliabilities for $X$ and $Y$ to differ across moderator-based subgroups. Sample sizes across moderator-based subgroups, which are related to the variance of the moderator variable, also are explicitly included in the model.\(^2\)

Theorem 2 also reveals that power decreases as the reliability of $X$ and/or $Y$ decreases. Furthermore, to a first-order approximation and if reliabilities are homogeneous across moderator-based subgroups, power depends on the reliabilities $\alpha_x$ and $\alpha_y$ only through their product $\alpha_x \alpha_y$. Accordingly, power is affected by measurement error in $X$ and $Y$ in a symmetric fashion. For example, holding constant all other factors affecting power, the power of MMR will be approximately the same for the case (a) $\alpha_x = .80$ and $\alpha_y = .80$ (i.e., $\alpha_x \alpha_y = .64$) as for the case (b) $\alpha_x = .90$ and $\alpha_y = .71$ (i.e., $\alpha_x \alpha_y = .639$).

Also, Theorem 2 clarifies the effects of $X$ and $Y$ variance heterogeneity on power. More precisely, holding constant all other factors shown in the theorem and given a situation with two moderator-based subgroups, $\sigma^2_{x,1}$, $\sigma^2_{x,2}$, $\sigma^2_{y,1}$, and $\sigma^2_{y,2}$ affect power only through the ratios $\sigma^2_{y,1}/\sigma^2_{y,2}$ and $\sigma^2_{x,1}/\sigma^2_{x,2}$. This effect is illustrated later in the article in the section Relative Impact of Factors Affecting Power.

Finally, Theorem 2 also provides theoretical evidence that complements previous empirical results regarding the interactive effects of various factors that affect the power of MMR (Aguinis & Stone-Romero, 1997; Stone-Romero et al., 1994). The presence of interactive effects on power suggests that even if the value regarding one
factor (e.g., total sample size) is favorable in terms of power (i.e., large), the existence of at least one other factor with an unfavorable value (e.g., poor reliability for \( X \)) may reduce power substantially. Thus, an unfavorable condition regarding any of the factors known to affect the power of MMR imposes a ceiling for the power of the \( F \) test. These interactive effects explain the typical concerns regarding the low power of MMR and the all-too-frequent failures to find so-called elusive moderating effects (Zedeck, 1971).

**Accuracy of the Power Approximation**

We conducted a Monte Carlo simulation to evaluate the power approximation shown in Theorem 2. The simulation was not intended to examine thoroughly a large number of conditions regarding the factors known to affect power (e.g., sample/population variances, reliabilities, sample size). Rather, our goal was to assess the accuracy of Theorem 2. Thus, we examined a set of diverse conditions considered to be typical in organization studies rather than a full factorial including all possible combinations of independent variable values included in the design. We implemented the simulation using a MATLAB 5 (http://www.mathworks.com/products/matlab) program.

**Independent Variables**

The simulation manipulated the following variables: (a) number of moderator-based subpopulations, (b) total sample size, (c) sample size across the moderator-based subgroups, (d) true score correlations between \( X \) and \( Y \) for the moderator-based subpopulations, (e) sampling restriction on \( X \), (f) variances of \( X \) and \( Y \), (g) reliabilities of \( X \) and \( Y \), and (h) deviation from normality for \( X \). Table 1 shows the 26 combinations of independent variable values, or cases, included in the simulation. Each of the 26 cases was sampled 10,000 times.

Commentary is needed regarding the manipulated variables. First, we chose to manipulate correlations as opposed to slopes. Although Appendix C shows that the moderating effect is defined as differences in slopes across the values of the moderator \( Z \) and the null hypothesis regarding the presence of the moderating effect is based on differences in slopes, the simulation varied correlations so as to make the results more easily interpretable. In other words, the standardized correlation metric might be more familiar and easier to interpret by most researchers than the unstandardized regression coefficient metric.

Second, regarding the \( X \) distribution, we sampled \( X \) scores from either a normal or a beta (1.5, 3.0) distribution. This beta distribution has a skewness coefficient of .51 and a kurtosis coefficient of −.46.

Third, regarding sampling restriction on \( X \), we implemented the following four types of restriction on the normal and beta distributions: (a) left truncation (i.e., scores are sampled if they are above a specific cutoff point), (b) right truncation (i.e., scores are sampled only if they are below a specific cutoff point), (c) sparse left (i.e., low scores are sparsely sampled, and high scores are more densely sampled), and (d) sparse right (i.e., high scores are sparsely sampled, and low scores are more densely sampled). These four types of sampling restrictions are defined mathematically in Appendix C. Each form of sampling restriction was crossed with four values of a truncation-like parameter (i.e., \( T = .00, .25, .50, \) and \( .75 \)), which is the proportion of
<table>
<thead>
<tr>
<th>Case Number</th>
<th>( k )</th>
<th>( n_j )</th>
<th>( \rho_{j} )</th>
<th>( T )</th>
<th>( \sigma_{x,j} )</th>
<th>( \sigma_{y,j} )</th>
<th>( \alpha_{x,j} )</th>
<th>( \alpha_{y,j} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>50, 50</td>
<td>.10, .30</td>
<td>.75, .75</td>
<td>.40, .40</td>
<td>.40, .40</td>
<td>.70, .70</td>
<td>.70, .70</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>30, 70</td>
<td>.10, .50</td>
<td>.75, .50</td>
<td>.40, 1.2</td>
<td>.40, 2.0</td>
<td>.70, 90</td>
<td>.70, 90</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>10, 90</td>
<td>.10, .70</td>
<td>.75, .25</td>
<td>.40, 2.0</td>
<td>.40, 1.2</td>
<td>.90, 70</td>
<td>.90, 70</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>125, 125</td>
<td>.30, 10</td>
<td>.50, .75</td>
<td>1.2, .40</td>
<td>2.0, .40</td>
<td>.90, 90</td>
<td>.90, 90</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>75, 175</td>
<td>.30, 50</td>
<td>.50, 50</td>
<td>1.2, 1.2</td>
<td>1.2, 2.0</td>
<td>.70, 70</td>
<td>.70, 70</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>25, 225</td>
<td>.30, 70</td>
<td>.50, .25</td>
<td>1.2, 2.0</td>
<td>2.0, 1.2</td>
<td>.70, 90</td>
<td>.70, 90</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>200, 200</td>
<td>.50, .10</td>
<td>.25, .75</td>
<td>2.0, .40</td>
<td>1.2, .40</td>
<td>.90, 70</td>
<td>.90, 70</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>120, 280</td>
<td>.50, .30</td>
<td>.25, .50</td>
<td>2.0, 1.2</td>
<td>2.0, .40</td>
<td>.90, 90</td>
<td>.90, 90</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>40, 360</td>
<td>.50, .70</td>
<td>.25, .25</td>
<td>2.0, 2.0</td>
<td>2.0, 1.2</td>
<td>.70, 90</td>
<td>.70, 90</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>125, 125</td>
<td>.10, .50</td>
<td>0</td>
<td>1.2, 1.2</td>
<td>2.0, 2.0</td>
<td>.90, 90</td>
<td>.90, 90</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>125, 125</td>
<td>.30, 30</td>
<td>0</td>
<td>1.2, 1.2</td>
<td>2.0, 2.0</td>
<td>.90, 90</td>
<td>.90, 90</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>50, 50</td>
<td>.10, .30</td>
<td>.75, .75</td>
<td>.40, .40</td>
<td>.40, .40</td>
<td>.70, 70</td>
<td>.70, 70</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>75, 175</td>
<td>.30, 50</td>
<td>.50, 50</td>
<td>1.2, 1.2</td>
<td>1.2, 2.0</td>
<td>.70, 70</td>
<td>.70, 70</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>25, 25, 50</td>
<td>.10, 30, 50</td>
<td>.75, 75, 75</td>
<td>.40, 40, 12</td>
<td>.40, 12, 20</td>
<td>.70, 70, 70</td>
<td>.70, 70, 70</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>30, 30, 40</td>
<td>.10, 50, .70</td>
<td>.75, 75, 50</td>
<td>.40, 12, 20</td>
<td>.40, 40, 12</td>
<td>.70, 90, 70</td>
<td>.70, 90, 70</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>35, 35, 30</td>
<td>.30, 10, .50</td>
<td>.75, 50, .25</td>
<td>.40, 20, 2.0</td>
<td>1.2, 20, .40</td>
<td>.70, 90, 90</td>
<td>.70, 90, 90</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>25, 75, 150</td>
<td>.30, 30, .70</td>
<td>.50, 75, .25</td>
<td>.40, 12, 1.2</td>
<td>2.0, 12, 4.0</td>
<td>.90, 70, 70</td>
<td>.90, 70, 70</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>50, 75, 125</td>
<td>.50, 10, .70</td>
<td>.50, 50, .25</td>
<td>1.2, 40, 2.0</td>
<td>40, 12, 1.2</td>
<td>.90, 90, 90</td>
<td>.90, 90, 90</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>75, 75, 100</td>
<td>.50, 10, .30</td>
<td>.50, 25, 25</td>
<td>1.2, 40, 1.2</td>
<td>1.2, 40, 2.0</td>
<td>.90, 90, 70</td>
<td>.90, 90, 70</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>50, 75, 275</td>
<td>.70, 30, 50</td>
<td>.25, 75, 50</td>
<td>1.2, 20, 4.0</td>
<td>2.0, 20, 4.0</td>
<td>.70, 70, 90</td>
<td>.70, 90, 90</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>75, 75, 250</td>
<td>.70, 50, 10</td>
<td>.25, 50, 25</td>
<td>2.0, 20, 4.0</td>
<td>1.2, 40, 2.0</td>
<td>.70, 90, 90</td>
<td>.70, 90, 90</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>90, 90, 220</td>
<td>.10, 70, .50</td>
<td>.25, 50, .75</td>
<td>1.2, 12, .40</td>
<td>2.0, 12, .40</td>
<td>.90, 70, 70</td>
<td>.90, 70, 70</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>75, 75, 100</td>
<td>.10, 30, .50</td>
<td>0, 0</td>
<td>1.2, 12, 1.2</td>
<td>2.0, 20, 2.0</td>
<td>.90, 90, 90</td>
<td>.90, 90, 90</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>75, 75, 100</td>
<td>.30, 30, 30</td>
<td>0, 0</td>
<td>1.2, 12, 1.2</td>
<td>2.0, 20, 2.0</td>
<td>.90, 90, 90</td>
<td>.90, 90, 90</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>25, 25, 50</td>
<td>.10, .30, .50</td>
<td>.75, 75, .75</td>
<td>.40, 40, 12</td>
<td>.40, 12, 20</td>
<td>.70, 70, 70</td>
<td>.70, 70, 70</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>75, 75, 100</td>
<td>.10, .30, .30</td>
<td>.50, 25, 25</td>
<td>2.0, 40, 1.2</td>
<td>1.2, 12, 2.0</td>
<td>.90, 90, 70</td>
<td>.90, 90, 70</td>
</tr>
</tbody>
</table>

**Note.** \( k \) = number of moderator-based subpopulations; \( n_j \) = sample size in each moderator-based subgroup (i.e., total sample size = \( \Sigma n_j \)); \( \rho_{j} \) = true score correlations between \( X \) and \( Y \) for each moderator-based subpopulation; \( T \) = truncation proportion on \( X \) for each moderator-based subgroup (i.e., proportion of scores that cannot be included in the sample); \( \sigma_{x,j} \) = true score standard deviation for \( X \) for each moderator-based subpopulation; \( \sigma_{y,j} \) = true score standard deviation for \( Y \) for each moderator-based subpopulation; \( \alpha_{x,j} \) = reliability for \( X \); \( \alpha_{y,j} \) = reliability for \( Y \) for each moderator-based subpopulation.
the population that cannot be sampled due to restriction. Thus, the manipulation of the sampling mechanism for $X$ led to various degrees of skewness, kurtosis, and $X$ variance multiplying factors (i.e., expected sample variance of $X$/population variance of $X$). Tables showing the values for skewness, kurtosis, and variance multiplying factors associated with each combination of sampling mechanism and $T$ for the normal and the beta (i.e., nonnormal) distributions are available from the authors on request.

Independent variable values. Commentary is needed regarding the choice for the various independent variable values used in the simulation and shown in Table 1. Once again, the goal of this simulation was to assess the accuracy of the analytically derived solution to approximate power. Thus, we were not attempting to include every situation encountered by researchers or to generate extensive power tables. Investigators can generate their own custom power values using the computer program that implements the analytic solution (see the Computer Program section below).

We chose the independent variable values according to the following rationale. First, regarding the number of moderator-based subpopulations, we chose the values of $k = 2$ and 3. This was based on a literature review by Aguinis et al. (2000) concluding that virtually all of the 616 MMR tests of categorical moderator variables reported in major organization studies journals over the past 30 years included moderators with two or three levels.

Second, regarding total sample size, we chose values ranging from 100 to 400. This choice was based on Jaccard and Wan’s (1995) review of American Psychological Association (APA) journals indicating that the median sample size is 175.

Third, regarding sample size across the moderator-based subgroups, we divided the total sample size into two or three subgroups based on proportions ranging from .10 to .50. The rationale was that researchers might find situations in which the proportion of scores in one subgroup is as low as .10 of the total sample. On the other hand, there might be situations in which the sample size is fairly equal across moderator-based subgroups (i.e., a proportion of .50). Thus, a range of .10 to .50 covers most typical research situations.

Fourth, regarding the magnitude of the moderating effect, we chose various combinations of correlations ranging from .10 to .70 across moderator-based subpopulations. The rationale was that these are the values typically observed, although .70 is less usual, in organization studies and other social science fields (e.g., political science, psychology). For example, Cohen (1988) defined effects of .10, .30, and .50 as small, medium, and large, respectively. We also included negative correlations within this range (i.e., Cases 12, 13, 25, and 26 in Table 1). Note, however, that the corresponding population correlations based on observable scores may differ from population correlations based on true scores because correlations between observable scores are affected by measurement error.

Fifth, regarding the truncation proportion on $X$ (i.e., $T$), we chose values of .00, .25, .50, and .75. The rationale was that these values cover a range from $T = 0$, in which all population scores can be included in the samples, to $T = .75$, in which only 25% of the population scores can be included in the sample. Implementing truncation on the normal and beta distributions led to variance multiplying factors (i.e., expected sample variance of $X$/population variance of $X$) ranging from 1.00 to 1.7. Thus, the simulation included situations ranging from no differences between sample and population variances to situations in which there is very severe variance restriction.
Sixth, the values for the standard deviation of X and Y were chosen after conducting a cursory review of several APA journals. Our review suggested that published studies using Likert-type scales with five and seven anchors yield standard deviations in the .40 to 2.0 range. Thus, this guided our choice for the standard deviation values.

Finally, regarding reliabilities for X and Y, we chose the values .70 and .90. Similar to Jaccard and Wan (1995), the rationale was that .70 is considered to be the lower bound of acceptable reliability levels (Nunnally & Bernstein, 1994), and .90 is a desirable level that serves as a de facto upper-bound level in many areas of organization studies and other social sciences.

**Dependent Variable**

The dependent variable, power, was the proportion of times out of each set of 10,000 trials that the null hypothesis of no moderating effect was rejected. We computed this proportion for each of the 26 cases shown in Table 1 crossed with each of the four sampling mechanisms (i.e., left and right truncation and sparse left and right) and the two types of underlying X distributions (i.e., normal and beta) described above and defined mathematically in Appendix C. We also computed an approximate (i.e., analytically derived) power value using Theorem 2 as implemented by the computer program MMRPOWER (see Computer Program section below).

**Results and Discussion**

Tables 2 and 3 show (a) the proportion of times the null hypothesis was empirically rejected, and (b) the power approximation yielded by Theorem 2 for each of the 26 cases shown in Table 1 crossed with the four sampling mechanisms. Table 2 shows results pertaining to an underlying normal X distribution, and Table 3 shows results for an underlying beta (i.e., nonnormal) X distribution. Tables 2 and 3 indicate that the empirical and theoretical proportions are virtually identical for every condition. The difference between the analytically derived power approximation and the empirically derived rejection rates was in no case greater than .019. The mean absolute deviation between the simulation and theory-based power values is .0036 for Table 2 and .0037 for Table 3. Given the statistical power metric (i.e., ranging from 0 to 1.00), the differences between the empirical and analytic values are negligible.

**Comparison with previous empirically based algorithms to estimate power.** We next compared analytically derived power estimates with those generated by the Aguinis and Pierce (1998b) program. Recall that, in contrast to the present analytic solution, the Aguinis and Pierce program suffers from certain limitations (e.g., does not allow for the consideration of measurement error, only allows for power estimation for binary moderator variables). Because of these limitations, the power estimates generated using the Aguinis and Pierce program assume that (a) reliabilities are 1.0 for X and Y across the two moderator-based subgroups, (b) the variance and truncation for X are identical across the two moderator-based subgroups, (c) the variance for Y is identical across the two moderator-based subgroups, and (d) X and Y scores follow a bivariate normal distribution. Because the Aguinis and Pierce program does not allow for the specification of as many variables as the present analytic approximation, in the comparison we could only vary (a) total sample size, (b) sample sizes across moderator-
based subgroups, (c) correlations between $X$ and $Y$ for each moderator-based subpopulation, and (d) truncation proportion. Table 4 shows results of this comparison.

Table 4 shows that, as was shown in Tables 2 and 3, the analytic approximation is virtually identical to the simulation-based results. In addition, Table 4 shows that, under restrictive assumptions such as lack of measurement error across subpopulations, the Aguinis and Pierce (1998b) program power estimate also was close to the simulation-based value in several conditions. For instance, for Cases 1 through 3, the difference between the simulation and the Aguinis and Pierce power values ranged from $-0.049$ to $0.021$. Note, however, that these cases include values that fall within the range of the simulation study on which the Aguinis and Pierce program was based (i.e., Aguinis & Stone-Romero, 1997). Thus, it was expected that the Aguinis and Pierce power estimate would be fairly accurate for these situations.

<table>
<thead>
<tr>
<th>Case Number</th>
<th>Left Truncation</th>
<th>Right Truncation</th>
<th>Theoretical Approximation</th>
<th>Sparse Left Approximation</th>
<th>Sparse Right Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.061</td>
<td>.060</td>
<td>.063</td>
<td>.077</td>
<td>.071</td>
</tr>
<tr>
<td>2</td>
<td>.000</td>
<td>.001</td>
<td>.000</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>3</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>.002</td>
<td>.002</td>
<td>.001</td>
<td>.008</td>
<td>.007</td>
</tr>
<tr>
<td>5</td>
<td>.169</td>
<td>.166</td>
<td>.161</td>
<td>.313</td>
<td>.311</td>
</tr>
<tr>
<td>6</td>
<td>.307</td>
<td>.309</td>
<td>.317</td>
<td>.313</td>
<td>.305</td>
</tr>
<tr>
<td>7</td>
<td>.021</td>
<td>.021</td>
<td>.019</td>
<td>.070</td>
<td>.066</td>
</tr>
<tr>
<td>8</td>
<td>.898</td>
<td>.896</td>
<td>.898</td>
<td>.984</td>
<td>.983</td>
</tr>
<tr>
<td>9</td>
<td>.295</td>
<td>.295</td>
<td>.288</td>
<td>.278</td>
<td>.295</td>
</tr>
<tr>
<td>10</td>
<td>.840</td>
<td>.840</td>
<td>.842</td>
<td>.846</td>
<td>.841</td>
</tr>
<tr>
<td>11</td>
<td>.048</td>
<td>.048</td>
<td>.050</td>
<td>.046</td>
<td>.048</td>
</tr>
<tr>
<td>12</td>
<td>.058</td>
<td>.063</td>
<td>.063</td>
<td>.080</td>
<td>.073</td>
</tr>
<tr>
<td>13</td>
<td>.165</td>
<td>.162</td>
<td>.161</td>
<td>.308</td>
<td>.319</td>
</tr>
<tr>
<td>14</td>
<td>.007</td>
<td>.005</td>
<td>.005</td>
<td>.007</td>
<td>.007</td>
</tr>
<tr>
<td>15</td>
<td>.010</td>
<td>.008</td>
<td>.007</td>
<td>.021</td>
<td>.019</td>
</tr>
<tr>
<td>16</td>
<td>.070</td>
<td>.069</td>
<td>.066</td>
<td>.082</td>
<td>.081</td>
</tr>
<tr>
<td>17</td>
<td>.423</td>
<td>.420</td>
<td>.411</td>
<td>.498</td>
<td>.500</td>
</tr>
<tr>
<td>18</td>
<td>.121</td>
<td>.122</td>
<td>.113</td>
<td>.220</td>
<td>.214</td>
</tr>
<tr>
<td>19</td>
<td>.303</td>
<td>.300</td>
<td>.301</td>
<td>.473</td>
<td>.468</td>
</tr>
<tr>
<td>20</td>
<td>.602</td>
<td>.597</td>
<td>.594</td>
<td>.782</td>
<td>.773</td>
</tr>
<tr>
<td>21</td>
<td>.101</td>
<td>.100</td>
<td>.102</td>
<td>.181</td>
<td>.189</td>
</tr>
<tr>
<td>22</td>
<td>.312</td>
<td>.311</td>
<td>.308</td>
<td>.448</td>
<td>.445</td>
</tr>
<tr>
<td>23</td>
<td>.586</td>
<td>.586</td>
<td>.583</td>
<td>.585</td>
<td>.591</td>
</tr>
<tr>
<td>24</td>
<td>.052</td>
<td>.050</td>
<td>.050</td>
<td>.053</td>
<td>.050</td>
</tr>
<tr>
<td>25</td>
<td>.006</td>
<td>.006</td>
<td>.005</td>
<td>.005</td>
<td>.006</td>
</tr>
<tr>
<td>26</td>
<td>.299</td>
<td>.301</td>
<td>.301</td>
<td>.472</td>
<td>.464</td>
</tr>
</tbody>
</table>

Note. Cases are defined in Table 1. Left truncation and right truncation = empirical power values (i.e., proportion of times the null hypothesis of no moderating effect was empirically rejected) implementing left truncation and right truncation sampling mechanisms defined in Appendix C; theoretical approximation = power approximate using Theorem 2 as implemented by the program MMRPOWER; sparse left and sparse right = empirical power values implementing sparse left and sparse right sampling restrictions mechanisms defined in Appendix C.
Alternatively, Cases 4 and 5 show situations including values falling outside of the range of the simulation work on which the Aguinis and Pierce (1998b) program was based (i.e., for Case 4, there is a negative correlation, and for Case 5, the correlations are greater than .80). In these situations, the difference between the simulation and the Aguinis and Pierce power values was greater. In Case 4, the Aguinis and Pierce program yielded a negatively biased value (i.e., −.170), and in Case 5, the program yielded a positively biased value (i.e., .789). These discrepancies illustrate one of the weaknesses described in the introduction section regarding the available computer programs to estimate power: Because they are based on Monte Carlo data, their accuracy is confined only to values similar to those used in the simulation on which the programs are based. The Aguinis and Pierce program is based on the Aguinis and Stone-Romero (1997) simulation, which did not include correlations lower than .20 or

Table 3
Comparison of Empirical and Analytic (i.e., using Theorem 2)
Power Values (underlying beta [1.5, 3.0] distribution for $X$)

<table>
<thead>
<tr>
<th>Case Number</th>
<th>Left Truncation</th>
<th>Right Truncation</th>
<th>Sparse Left</th>
<th>Sparse Right</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation</td>
<td>Theory</td>
<td>Simulation</td>
<td>Theory</td>
</tr>
<tr>
<td>1</td>
<td>.064</td>
<td>.065</td>
<td>.051</td>
<td>.053</td>
</tr>
<tr>
<td>2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>.225</td>
<td>.217</td>
<td>.080</td>
<td>.077</td>
</tr>
<tr>
<td>5</td>
<td>.308</td>
<td>.318</td>
<td>.308</td>
<td>.319</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.958</td>
<td>.958</td>
<td>.695</td>
<td>.690</td>
</tr>
<tr>
<td>8</td>
<td>.297</td>
<td>.290</td>
<td>.295</td>
<td>.287</td>
</tr>
<tr>
<td>9</td>
<td>.841</td>
<td>.842</td>
<td>.841</td>
<td>.842</td>
</tr>
<tr>
<td>10</td>
<td>.045</td>
<td>.050</td>
<td>.050</td>
<td>.050</td>
</tr>
<tr>
<td>11</td>
<td>.064</td>
<td>.065</td>
<td>.053</td>
<td>.053</td>
</tr>
<tr>
<td>12</td>
<td>.219</td>
<td>.217</td>
<td>.077</td>
<td>.077</td>
</tr>
<tr>
<td>13</td>
<td>.007</td>
<td>.005</td>
<td>.005</td>
<td>.004</td>
</tr>
<tr>
<td>14</td>
<td>.008</td>
<td>.006</td>
<td>.001</td>
<td>.001</td>
</tr>
<tr>
<td>15</td>
<td>.072</td>
<td>.068</td>
<td>.069</td>
<td>.070</td>
</tr>
<tr>
<td>16</td>
<td>.456</td>
<td>.444</td>
<td>.381</td>
<td>.372</td>
</tr>
<tr>
<td>17</td>
<td>.155</td>
<td>.143</td>
<td>.083</td>
<td>.078</td>
</tr>
<tr>
<td>18</td>
<td>.382</td>
<td>.381</td>
<td>.211</td>
<td>.217</td>
</tr>
<tr>
<td>19</td>
<td>.675</td>
<td>.667</td>
<td>.384</td>
<td>.378</td>
</tr>
<tr>
<td>20</td>
<td>.134</td>
<td>.133</td>
<td>.064</td>
<td>.067</td>
</tr>
<tr>
<td>21</td>
<td>.380</td>
<td>.371</td>
<td>.198</td>
<td>.196</td>
</tr>
<tr>
<td>22</td>
<td>.590</td>
<td>.583</td>
<td>.590</td>
<td>.583</td>
</tr>
<tr>
<td>23</td>
<td>.048</td>
<td>.050</td>
<td>.048</td>
<td>.050</td>
</tr>
<tr>
<td>24</td>
<td>.006</td>
<td>.005</td>
<td>.006</td>
<td>.004</td>
</tr>
<tr>
<td>25</td>
<td>.392</td>
<td>.381</td>
<td>.220</td>
<td>.217</td>
</tr>
</tbody>
</table>

Note. Cases are defined in Table 1. Left and right truncation = sampling mechanisms based on truncation defined in Appendix C; sparse left and sparse right = sparse left and sparse right sampling restriction mechanisms defined in Appendix C; simulation = empirical power values (i.e., proportion of times the null hypothesis of no moderating effect was empirically rejected); theory = power approximate using Theorem 2 as implemented by the program MMRPOWER.
higher than .80. The foregoing example illustrates the superiority of an analytic solution that can be generalized to any parameter value range. It is also of interest to approximate power for the cases shown in Table 4 once some of the Aguinis and Pierce (1998b) assumptions are relaxed. For instance, Case 1 yielded a satisfactory power value (.961 for the Aguinis & Pierce program and .940 for the simulation; a difference of just .021). Recall that these power values assume perfect reliability for both $X$ and $Y$ scores for the two moderator-based subgroups. Relaxing this assumption, we generated a power value via simulation assuming a more realistic scenario in organization studies in which reliabilities for $X$ and $Y$ are .80 for each of the two moderator-based subgroups. The resulting simulation power value was .734, below the recommended value of .80 (Cohen, 1988). Thus, just by relaxing one of the assumptions (i.e., lack of measurement error), the difference between the Aguinis and Pierce and the simulation power values increased from .021 to .229. In contrast, the power value generated using the analytic approximation was .736, a difference of just .002 from the simulation result. This example shows a second limitation mentioned in the introduction section pertaining to previous research attempting to estimate the power of MMR based on empirical work. Once again, computer programs based on empirically derived algorithms to estimate power (e.g., Aguinis et al., 1994; Aguinis & Pierce, 1998b) cannot go beyond the empirical work on which they are based; for instance, if the original research assumed lack of measurement error, so do the corresponding computer programs. In contrast, the present analytic solution is based on theory and goes beyond previous empirical work. Consequently, it allows power values to be approximated based on more realistic conditions (e.g., measurement error for $X$ and $Y$ across moderator-based subgroups, nonnormal $X$ distributions, range restriction beyond simple truncation on $X$). In short, researchers using the present analytic result have a more generalizable and accurate approximation of power values.

Table 4

<table>
<thead>
<tr>
<th>Case Number</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$\rho_j$</th>
<th>$\rho_j$</th>
<th>$T$</th>
<th>Analytic Approximation</th>
<th>A&amp;P</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>150, 150</td>
<td>.20, .80</td>
<td>.60</td>
<td>.941</td>
<td>.961</td>
<td>.940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>50, 100</td>
<td>.20, .40</td>
<td>.00</td>
<td>.228</td>
<td>.184</td>
<td>.233</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>80, 120</td>
<td>.10, .40</td>
<td>.90</td>
<td>.147</td>
<td>.158</td>
<td>.145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>75, 100</td>
<td>-.40, .40</td>
<td>.20</td>
<td>.989</td>
<td>.816</td>
<td>.986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>150, 150</td>
<td>.95, .99</td>
<td>.60</td>
<td>.124</td>
<td>.910</td>
<td>.121</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $n_j$ = sample size in each moderator-based subgroup (i.e., total sample size = $n_1 + n_2$); $\rho_j$ = correlations between $X$ and $Y$ for each moderator-based subpopulation; $T$ = truncation proportion on $X$ for each moderator-based subgroup (i.e., proportion of scores that cannot be included in the sample); analytic approximation = power approximate using Theorem 2 as implemented by the program MMRPOWER; A&P = power value generated using the Aguinis and Pierce (1998b) computer program; simulation = empirical power values (i.e., proportion of times the null hypothesis of no moderating effect was empirically rejected). The comparison in this table holds $k$ constant at 2 (i.e., the moderator variable has two levels), reliabilities of $X$ and $Y$ constant at 1.00, and variance of $Y$ constant at 1.00 (i.e., the Aguinis & Pierce, 1998b, program does not allow for variations in $k$, reliabilities, and $Y$ variance).
and, consequently, are likely to make more informed conclusions regarding the operation of moderating effects of categorical variables.

**Computer Program**

As noted above, we developed a computer program that performs all the necessary computations required by the theoretical approximation shown in Theorem 2. This program (MMRPOWER) is available at http://www.math.montana.edu/~rjboik/power.html. We chose to make MMRPOWER available on the Internet so as to reach the largest possible number of users regardless of operating system platform (e.g., Windows 95/98/NT/2000, Macintosh, OS2). The program was written in FORTRAN 77 and consists of a main program that calls several FORTRAN subprograms and functions that were originally published in *Applied Statistics* (a list of the algorithms used is available from the authors on request).

The first screen of the program prompts the user to provide information regarding (a) number of moderator-based subgroups (the maximum number handled by the program is 20), (b) desired significance level or test size (i.e., preset Type I error rate), (c) desired test (i.e., overall test of equality of slopes across moderator-based subgroups or tests of specific contrasts of slopes), and (d) sampling restrictions (i.e., none, sampling from truncated \(X\) normal distributions, or sampling from nonnormal \(X\) distributions). The program also prompts the user to provide information regarding whether the input format includes correlations based on true scores, correlations based on observable scores, slopes based on true scores, or slopes based on observable scores.

The second screen of the program prompts the user to input the necessary information to compute power. In addition to sample size and reliabilities for each of the moderator-based subgroups, the necessary input varies depending on the choices made on the first screen. That is, the user is prompted for correlations or slopes based on true or observable scores for each moderator-based subgroup. In addition, the user is prompted for the truncation proportion for \(X\) (i.e., \(T\)) if truncated normal distributions were noted on the first screen, no information if no sampling restrictions were noted on the first screen, and variance multiplying factors (i.e., expected sample variance/population variance) if nonnormal \(X\) distributions were noted on the first screen. Finally, if tests of specific contrasts were requested on the first screen, the second screen prompts the user for contrast coefficients (see Boik, 1979, 1993, for a discussion of main and interaction effect contrasts).

**Needed Input**

All the information required by the program is typically available to researchers (except for truncation and variance multiplying factor information). If the true score options are chosen, sample-based statistics must be used to estimate parameters. The program will yield an accurate power value as long as the estimates are accurate.

Information regarding truncation and the variance multiplying factor (when the sampling mechanism is other than truncation) may not be available. If this information is not available, we suggest that researchers use an estimate based on relevant literature. If a literature-based estimate is not available, researchers can input a best-case scenario (i.e., use a truncation proportion of 0.00 and assume no truncation or a variance multiplying factor of 1.00 and assume no variance difference between the sample
and the population) and a worse-case scenario (i.e., use a severe truncation proportion such as 0.75 or a severe variance multiplying factor of 0.25). Given the absence of information on truncation and the variance multiplying factor, researchers will know that the power of their MMR test lies somewhere between the best-case and worse-case situations.

Relative Impact of Factors Affecting Power

As noted in the introduction section, several empirical studies have examined the impact of each of the many factors known to affect the power of MMR. Typically, these Monte Carlo simulations have investigated only the concurrent impact of two or three factors. For instance, Stone-Romero et al. (1994) only manipulated three design-related factors: (a) sample size in each of two moderator variable-based subgroups, (b) total sample size, and (c) magnitude of moderating effect. Likewise, Stone-Romero and Anderson (1994) also varied only three factors: (a) total sample size, (b) unreliability of predictor variable scores, and (c) magnitude of moderating effect. Another contribution of the present analytic approximation is that, in contrast to previous empirical work, power values can be easily and concurrently generated for a diverse set of variables and values. Consequently, we can now gain a better understanding of, when other variables are held constant, what is the relative impact of improving one factor (e.g., increasing total sample size) as compared to improving conditions regarding another factor (e.g., minimizing restriction on X).

We present a total of 34 illustrative cases in Tables 5 and 6 to show the relative impact of the various factors affecting the power of MMR. In Table 5, we illustrate the effects of total sample size, sample size across moderator-based subgroups, differences in correlations across moderator-based subgroups, truncation (for the X normal distribution case) and variance multiplying factor (for the arbitrary X distribution case), and reliabilities on X and Y.

Regarding Table 5, to make the various comparisons easier to understand, we present a situation in which there are two moderator-based subgroups (e.g., gender). Also, truncation, variance multiplying factor, and X and Y variances (σ_x = σ_y = 1.0) are not varied across the subpopulations (effects of X and Y variance heterogeneity are shown in Table 6). In addition, because the effects of reliability of X and Y are symmetrical when reliabilities are homogeneous across moderator-based subgroups, we varied the product α_xα_y, and not the reliability for each variable. Finally, we also held α_xα_y constant across the two subgroups.

Case 1 in Table 5 is what could be labeled an “optimal” situation for detecting a moderating effect. Total sample size is 400 (much larger than the median of 175 reported by Jaccard & Wan, 1995), sample size is equal across the subgroups, there is a .50 difference between the X – Y correlations across the two subgroups, there is no restriction on X, X and Y variances are equal across the two subgroups, and the product of the reliability terms is .81 (i.e., α_x = .90 and α_y = .90). Not surprisingly, given these optimal conditions, the power of the MMR test is .998.

Case 6 is what could be labeled an “average” or more typical situation for detecting a moderating effect. N = 175, the sample sizes ratio across the subgroups is .67 (i.e., they are dissimilar but not drastically different), there is a .30 difference between X – Y correlations across the two subgroups, there is some restriction (i.e., scores can be sampled from 75% of the range of population scores), variances of X and Y are identi-
cal across subpopulations, and the product of the reliability terms is .64 (i.e., α_y = α_x = .80 across subgroups). The resulting power for what can be considered a more typical research situation is .203. Given this low value, it is not surprising that moderating effects have been labeled “elusive” (Zedeck, 1971). If Case 6 indeed represents a situation faced by the majority of organization science researchers using MMR to test hypotheses regarding moderating effects of categorical variables, the chances of rejecting a null hypothesis would be greater if a coin toss were used instead of MMR.

Cases 2 through 5 in Table 5 are follow-ups to the optimal Case 1. We changed the value for each of the factors from optimal to average without altering the optimal conditions regarding all other factors. For instance, Case 2 shows that decreasing the total sample size and not having identical sizes across the subgroups decreases power from .998 to .867, holding all other factors at their optimal value. Case 3 shows that decreasing the difference in correlations from .10 in Subgroup 1 and .60 in Subgroup 2 to .10

### Table 5

<table>
<thead>
<tr>
<th>Case Number</th>
<th>nj</th>
<th>ρ_j</th>
<th>T</th>
<th>δ</th>
<th>x_y</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underlying normal X distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>200, 200</td>
<td>.10, .60</td>
<td>.00</td>
<td>.81</td>
<td>.998</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>105, 70</td>
<td>.10, .60</td>
<td>.00</td>
<td>.81</td>
<td>.867</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>200, 200</td>
<td>.10, .40</td>
<td>.00</td>
<td>.81</td>
<td>.791</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>200, 200</td>
<td>.10, .60</td>
<td>.25</td>
<td>.81</td>
<td>.942</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>200, 200</td>
<td>.10, .60</td>
<td>.00</td>
<td>.64</td>
<td>.988</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>105, 70</td>
<td>.10, .40</td>
<td>.25</td>
<td>.64</td>
<td>.203</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>200, 200</td>
<td>.10, .40</td>
<td>.25</td>
<td>.64</td>
<td>.431</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>105, 70</td>
<td>.10, .60</td>
<td>.25</td>
<td>.64</td>
<td>.494</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>105, 70</td>
<td>.10, .40</td>
<td>.00</td>
<td>.64</td>
<td>.341</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>105, 70</td>
<td>.10, .40</td>
<td>.25</td>
<td>.81</td>
<td>.248</td>
<td></td>
</tr>
<tr>
<td>Underlying arbitrary X distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>200, 200</td>
<td>.10, .60</td>
<td>1.00</td>
<td>.81</td>
<td>.998</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>105, 70</td>
<td>.10, .60</td>
<td>1.00</td>
<td>.81</td>
<td>.867</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>200, 200</td>
<td>.10, .40</td>
<td>1.00</td>
<td>.81</td>
<td>.791</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>200, 200</td>
<td>.10, .60</td>
<td>.75</td>
<td>.81</td>
<td>.987</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>200, 200</td>
<td>.10, .60</td>
<td>1.00</td>
<td>.64</td>
<td>.988</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>105, 70</td>
<td>.10, .40</td>
<td>0.75</td>
<td>.64</td>
<td>.268</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>200, 200</td>
<td>.10, .40</td>
<td>0.75</td>
<td>.64</td>
<td>.562</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>105, 70</td>
<td>.10, .60</td>
<td>0.75</td>
<td>.64</td>
<td>.638</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>105, 70</td>
<td>.10, .40</td>
<td>1.00</td>
<td>.64</td>
<td>.341</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>105, 70</td>
<td>.10, .40</td>
<td>0.75</td>
<td>.81</td>
<td>.329</td>
<td></td>
</tr>
</tbody>
</table>

Note. n_j = sample size in each moderator-based subgroup (i.e., total sample size = n_1 + n_2); ρ_j = correlations between X and Y for each moderator-based subpopulation; T = truncation proportion on X for each moderator-based subgroup (i.e., proportion of scores that cannot be included in the sample) for normal distribution; δ = variance multiplying factor (i.e., expected sample variance of X/population variance of X) for arbitrary distribution; α_y = reliability for Y and α_x = reliability for X; power = power approximate using Theorem 2 as implemented by the program MMRPOWER. Power values were approximated using a preset nominal Type I error = .05. It is assumed that σ_y (true score standard deviation for Y) = σ_x (true score standard deviation for X) = 1.00 for each moderator-based subpopulation.
and .40, respectively, decreases power from .998 to .791. Case 4 shows the effect of adding some restriction (i.e., scores can only be sampled from 75% of the population score range), and Case 5 shows the effect of adding more measurement error. Relatively speaking, and given the values shown in Cases 2 through 5 in Table 5, the following is a rank ordering of factors that affect the power of MMR in order of importance: (a) moderating effect magnitude, (b) total sample size, (c) sampling restriction on \(X\), and (d) measurement error. Of course, this rank ordering may change if noticeably different values are chosen as optimal and average for each of the factors.

Table 5 also shows various situations (Cases 7-10) in which there is an average condition regarding all factors, and there is improvement in one factor at a time, from average to optimal, without altering the other factors. As expected, the pattern of power improvement is similar to the pattern observed for Cases 2 through 5 regarding power decrements. Given that conditions are average for all other factors, increasing effect size and sample size, reducing sampling restriction on \(X\), and improving reliability increase power, in this respective order.

The bottom half of Table 5 shows a similar pattern of results when normality is not assumed for the \(X\) distribution in assessing the effects of differences between expected sample variance of \(X\) and population variance of \(X\). Effect size and sample size are the
two most important factors affecting power. However, the effects of reliability and restriction are similar in magnitude. That is, improving one or the other factor yields similar gains in power.

An additional conclusion can be drawn from Table 5. Aguinis and Stone-Romero (1997) concluded that the power function is nonlinear. Table 5 (and additional tables and graphs available from the authors on request) shows further that the pattern of nonlinearity depends on the factor that is varied as well as the magnitude of the change in the values of the factors manipulated.

Table 6 shows the effects on power of heterogeneity of $X$ and $Y$ variance across moderator-based subpopulations. As in Table 5, to make the various comparisons easier to understand, we present a situation in which there are two moderator-based subgroups (e.g., gender). Also, we assume no restriction on $X$ (i.e., $T = 0$ for the normal case and variance multiplying factor = 1.00 for the arbitrary distribution). In addition, because the effects of reliability of $X$ and $Y$ are symmetrical when $X$ and $Y$ reliabilities are homogeneous across moderator-based subgroups, we varied the product $\alpha_y \alpha_x$ and not the reliability for each variable. Finally, we also held $\alpha_y \alpha_x$ constant across the two subgroups.

Case 1 in Table 6 shows what could be labeled an optimal case for detecting a moderating effect. Sample size is large (i.e., 400) and equal across moderator-based subgroups, there is a .50 difference in the $X - Y$ correlations between the subpopulations, reliabilities for $X$ and $Y$ are .90 for each of the subgroups, and variances for $X$ and $Y$ are identical across the subpopulations. More important, because it is the variance ratios that have an impact on power (holding all other variables constant) rather than the absolute variance values, the ratios are $\sigma_y/\sigma_x$ = 1.00. Finally, the error variances are homogeneous, and their ratio is 1.40 (DeShon & Alexander, 1996, concluded that error variances are homogeneous if the ratio of the largest to the smallest variance is less than 1.50; see Aguinis & Pierce, 1998a, for a review of effects of error variance heterogeneity on the power of MMR). Not surprisingly, given these optimal, and perhaps rare, conditions, the power of MMR is .998.

In Cases 2 through 4 in Table 6, we systematically varied the $\sigma_y/\sigma_x$ and $\sigma_x/\sigma_y$ ratios to include values of 0.5, 1.0, and 2.0 and held all other variables identical to the optimal Case 1. Thus, in Case 2 the ratios are $\sigma_y/\sigma_x$ = 1.0 and $\sigma_x/\sigma_y$ = 0.5; in Case 3 the ratios are $\sigma_y/\sigma_x$ = 0.5 and $\sigma_x/\sigma_y$ = 1.0; and in Case 4 the ratios are $\sigma_y/\sigma_x$ = 0.5 and $\sigma_x/\sigma_y$ = 0.5. Note that as $X$ and $Y$ variances change, so do error variances and slopes. In Cases 5 through 7, we varied the $X$ and $Y$ variance ratios in an identical fashion as compared to Cases 2 through 4 but replaced each 0.5 ratio with a 2.0 ratio. Taken together, Cases 1 through 7 show that even when other conditions are optimal for the detection of a moderating effect with respect to reliability, sample size, and differences in correlations, heterogeneity of variance across subgroups can be detrimental to power. Some patterns of variance heterogeneity lead to a decrease in differences in slopes and, consequently, a decrease in power. For instance, Case 6 shows that despite the fact that $X$ variances are identical across subpopulations, a 2:1 ratio regarding $Y$ variance leads to a power of .645 as compared to a power of .998 when the $Y$ variances are identical. This effect is particularly noteworthy given that a 2:1 ratio of variances might not be perceived by most researchers as posing a serious threat to MMR-based conclusions. Nevertheless, the decrease in power is substantial.

Case 8 in Table 6 shows what could be labeled an average situation for detecting a moderating effect. Total sample size is 175, sample sizes are moderately different
across subgroups, $X - Y$ correlations are .10 and .40 across the two subpopulations, and reliabilities for $X$ and $Y$ are assumed to be .80 for each subgroup. Note, however, that in Case 8 the $X$ and $Y$ variances take on the optimal value such that they are identical across the subpopulations. Moreover, the ratio of error variances is also optimal. The ratio is 1.11, which is less than the 1.50 rule of thumb recommended to decide whether error variances are heterogeneous (DeShon & Alexander, 1996). This average situation yielded a power value of .341 (note that power is higher than the .203 value for the average situation in Case 6 in Table 5 because in Table 6 it is assumed that there is no restriction on $X$). Once again, as illustrated in Table 5, this shows that numerous, if not most, hypothesis tests using MMR are prone to Type II errors (i.e., incorrectly failing to reject a false null hypothesis).

Next, in Cases 9 through 14 in Table 6, we replicated the pattern of $X$ and $Y$ variance ratios used for Cases 2 through 7 to investigate the impact of the variance ratios on the power of this average situation. For instance, Case 10 shows that increasing $Y$ variance heterogeneity can increase power, even when there is error variance heterogeneity. Taken together, power values shown in Cases 9 through 14 reinforce a conclusion reached by DeShon and Alexander’s (1996) empirical study: Error variance can lead to increases or decreases in power because Type I error rates are not fixed at their preset nominal level (.05 in this case). Most notably, Cases 9 through 14 illustrate the impact of variance heterogeneity on slope differences and, in turn, on power. For instance, Cases 10 and 13 provide a clear example. Both cases are identical. The only difference is that for Case 10, $\sigma_{y,1}/\sigma_{y,2} = .5$, whereas for Case 13, $\sigma_{y,1}/\sigma_{y,2} = 2.0$. This difference in $Y$ variances led to power = .681 for Case 10 and power = .065 for Case 13. In short, what might be considered small differences in $Y$ variances across subpopulations can have dramatic effects on power.

There is an important point that should be emphasized regarding the effects of error variance heterogeneity in MMR. Results based on MMR cannot be trusted in the presence of error variance heterogeneity because Monte Carlo studies as well as the analytic approximation have demonstrated that MMR’s $F$ test does not control Type I error at the nominal rate when error variances are heterogeneous (Aguinis et al., 1999; Aguinis & Pierce, 1998a). Depending on population and sample characteristics, violating the homogeneity of error variance assumption (a) increases or decreases Type I error rates (e.g., a researcher mistakenly concludes that a moderating effect exists), and (b) increases or decreases Type II error rates (i.e., a researcher mistakenly concludes that a moderating effect does not exist). For instance, regarding Type I error, Dretzke, Levin, and Serlin (1982) showed that error rates are artificially inflated when sample sizes are unequal across subgroups, and this is most noticeable when the smaller subgroup sample size is paired with the larger error variance. Regarding Type II error, Alexander and DeShon (1994) found that error rates increase (i.e., statistical power is lowered) when the subgroup with the larger sample size is associated with the larger error variance (see Aguinis & Pierce, 1998a, for a review). Because of the Type I and Type II error rate problems due to heterogeneity of error variance, Aguinis et al. (1999) developed the computer program ALTMMR (available at http://members.aol.com/imsap/altmmr.html) to allow MMR users to (a) assess whether error variances are homogeneous, and (b) compute alternatives to MMR’s $F$ test if they are heterogeneous. Thus, we suggest that, after data have been collected, researchers use ALTMMR to check whether the homogeneity of error variance assumption is satisfied before proceeding to conduct an MMR analysis.
The present analytic solution to computing the power of MMR does not solve the problem that the \( F \) statistic does not have an \( F \) distribution when the homogeneity of error variance assumption is violated. MMR's results about whether \( H_0 \) should be rejected cannot be trusted when error variances are heterogeneous, and this is a problem associated with MMR and not with the present analytic solution. In fact, results show that power values generated using the analytic solution are virtually identical to those generated empirically, even when error variances are heterogeneous. The algorithm and program yield accurate power rates, but when the assumption is violated, these power rates correspond to a test whose size differs from the nominal \( \alpha \). In short, when error variances are heterogeneous, MMR results cannot be trusted, and MMR should not be used (Aguinis et al., 1999).

An additional contribution of the present analytic solution is that MMRPOWER can also be used to approximate the actual \( \alpha \), as well as learn about the deviation of the actual \( \alpha \) from the nominal \( \alpha \) given a violation of the homogeneity of error variance assumption. To do this, users would first select identical slopes for the moderator-based subgroups and then input various combinations of values for the factors known to cause a violation of the assumption (e.g., sample sizes, standard deviations).

MMRPOWER issues a caution statement in situations when \( H_0 \) is true and the actual \( \alpha \) differs from the preset nominal \( \alpha \), and MMRPOWER's output includes each group's error standard deviation (i.e., “W-Group Error SD”). The same caution statement is issued for any contrasts if the contrast \( H_0 \) is true and error variances are heterogeneous. In such situations, MMRPOWER outputs the following message:

The null hypothesis is true, but the actual significance level of the test is \( \alpha' \) [a numeric value is shown on the screen]. To obtain a new test with actual significance equal to 0.05000 set the nominal alpha to \( \alpha'' \) [a numeric value is shown on the screen]. The adjusted critical F value is \( F \) [a numeric value is shown on the screen].

The ease with which power values are generated using the present analytic solution as implemented by the program MMRPOWER also has practical research value. More precisely, generating power values for various hypothetical scenarios can aid researchers in making decisions about study design and allocating their research resources. For instance, assume a typical situation in which research resources are limited. Also, assume that a researcher is planning a study including a self-report survey and has a choice between two sets of measures for \( X \) and \( Y \). Based on the previous use of these measures, the first set would lead to \( \alpha_1 \alpha_2 = .81 \) (i.e., \( \alpha_1 = \alpha_2 = .90 \)), and these are paper-and-pencil instruments that take about 40 minutes to be completed. The second set of measures would lead to \( \alpha_1 \alpha_2 = .60 \) (i.e., \( \alpha_1 = .80 \) and \( \alpha_2 = .75 \)), and these instruments take approximately 20 minutes to be completed. Assume that past research shows that, because of the difference in time and effort involved in completing the measures, using the second set of measures improves response rate by about 15%. Should this researcher attempt to increase \( N \) by using the second set of measures or, alternatively, should this researcher use the first set of measures so as to have less measurement error in \( X \) and \( Y \) scores? Which of the two study scenarios will lead to greater power in estimating whether \( Z \) moderates the relationship between \( X \) and \( Y \)? What would be the resulting power value if \( \alpha_1 \alpha_2 \) decreases from .81 to .60 (i.e., by using the second set of measures) but total sample size increases from 200 to 230 (i.e., a 15% increase)? With
MMRPOWER, these questions can be answered via the analytic approximation to compute power for each of the anticipated situations.

**Conclusion**

MMR is a pervasively used statistical technique to estimate and test effects of categorical moderator variables in organization studies and the social sciences in general. Despite its popularity, researchers often express the concern that the power of MMR is inadequate to test hypotheses regarding the operation of moderating effects. Theorem 2 in Appendix C offers an analytic synthesis of factors affecting the power of MMR. Results of our Monte Carlo simulation indicate that Theorem 2 is accurate at approximating the power of MMR under a very diverse set of conditions including, among other factors, heterogeneity of variance of $X$, heterogeneity of variance of $Y$, and heterogeneity of reliabilities across moderator-based subgroups. We also developed the computer program MMRPOWER that implements the algorithm shown in Theorem 2. MMRPOWER can be used to study further the effects of violating certain assumptions on the power of MMR. Using MMRPOWER with values typically encountered in research situations in which MMR is used yielded power values substantially below Cohen’s (1988) recommended .80 value. Thus, it is likely that conclusions of many past studies that used MMR and did not find support for a hypothesized categorical moderator variable were actually the product of a Type II error (i.e., inability to correctly reject a false null hypothesis). In closing, we encourage researchers to approximate power while planning their research design (e.g., sample size) and methodology (e.g., reliability of measurement). By doing so, researchers will make more informed conclusions about hypotheses entailing moderating effects of categorical variables.

**APPENDIX A**

**Assumptions**

We consider two cases. In Case 1, the continuous criterion variable $Y$ and the continuous predictor $X$ are assumed to follow a bivariate normal distribution within each categorical moderator-based subpopulation. In Case 2, the conditional distribution of $Y$ given $X = x$ is assumed to be normal within each moderator-based subpopulation, but the marginal distributions of $X$ are arbitrary.

**True Scores and Error**

For Cases 1 and 2, it is assumed that the observable random variables $X$ and $Y$ can be decomposed as the sum of two statistically independent components, namely, true scores and error. That is,

$$
\begin{pmatrix}
Y \\
X
\end{pmatrix} = \begin{pmatrix}
Y_{\text{true}} \\
X_{\text{true}}
\end{pmatrix} + \begin{pmatrix}
\varepsilon_Y \\
\varepsilon_X
\end{pmatrix},
$$

(A1)

where $X_{\text{true}}$ and $Y_{\text{true}}$ are true scores and $\varepsilon_X$ and $\varepsilon_Y$ are random errors. Denote the $k$ subpopulations by $P_1, P_2, ..., P_k$. Then, the expectation and covariance matrix corresponding to the true scores in subpopulation $j$ can be written as
\[
E \left( \frac{Y_{\text{true}}}{X_{\text{true}}} \mid P_j \right) = \mu_j = \begin{pmatrix} \mu_{y,j} \\ \mu_{x,j} \end{pmatrix}
\quad \text{and}
\Var \left( \frac{Y_{\text{true}}}{X_{\text{true}}} \mid P_j \right) = \Sigma_{\text{true},j} = \begin{bmatrix} \sigma_{y,j}^2 & \rho_{j} \sigma_{y,j} \sigma_{x,j} \\ \rho_{j} \sigma_{y,j} \sigma_{x,j} & \sigma_{x,j}^2 \end{bmatrix},
\]

where the notation \( |P_j \) means that the result is specific to subpopulation \( j \); \( \rho_{j} \) is the correlation between \( X_{\text{true}} \) and \( Y_{\text{true}} \) in subpopulation \( j \); and \( \sigma_{y,j} \) and \( \sigma_{x,j} \) are the \( X_{\text{true}} \) and \( Y_{\text{true}} \) population standard deviations.

The random errors \( \varepsilon_y \) and \( \varepsilon_x \) are assumed to be distributed independently of one another. Specifically,

\[
\varepsilon_y \mid P_j \sim N(0, \sigma_{\varepsilon_y,j}^2) \quad \text{and} \quad \varepsilon_x \mid P_j \sim (0, \sigma_{\varepsilon_x,j}^2).
\]

Note that \( \varepsilon_y \) is assumed to be normally distributed, whereas only the first two moments, but not the distribution of \( \varepsilon_x \), is specified in Equation A3. In Case 1, it is assumed that \( \varepsilon_x \) is normally distributed, but in Case 2, the distribution of \( \varepsilon_x \) is arbitrary.

It follows from Equations A1, A2, and A3 that the vector of observable scores has the following expectation and covariance matrix:

\[
E \left( \frac{Y}{X} \mid P_j \right) = \mu_j \quad \text{and} \quad \Var \left( \frac{Y}{X} \mid P_j \right) = \Sigma_j = \begin{bmatrix} \sigma_{y,j}^2 + \sigma_{\varepsilon_y,j}^2 & \rho_{j} \sigma_{y,j} \sigma_{\varepsilon_x,j} \\ \rho_{j} \sigma_{y,j} \sigma_{\varepsilon_x,j} & \sigma_{x,j}^2 + \sigma_{\varepsilon_x,j}^2 \end{bmatrix}.
\]

Accordingly, the reliabilities of \( X \) and \( Y \) are

\[
\alpha_{x,j} = \frac{\sigma_{x,j}^2}{\sigma_{x,j}^2 + \sigma_{\varepsilon_x,j}^2} \quad \text{and} \quad \alpha_{y,j} = \frac{\sigma_{y,j}^2}{\sigma_{y,j}^2 + \sigma_{\varepsilon_y,j}^2},
\]

respectively.

**Case 1: Bivariate Normal Distribution for \((Y,X)\)**

In Case 1, it is assumed that the vector of true scores as well as the vector of random errors follow bivariate normal distributions. That is,

\[
\begin{bmatrix} Y_{\text{true}} \\ X_{\text{true}} \end{bmatrix} \mid P_j \sim N(\mu_j, \Sigma_{\text{true},j}) \quad \text{and} \quad \begin{bmatrix} \varepsilon_y \\ \varepsilon_x \end{bmatrix} \mid P_j \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\varepsilon_y,j}^2 & 0 \\ 0 & \sigma_{\varepsilon_x,j}^2 \end{bmatrix}\right)
\]

where \( \mu_j \) and \( \Sigma_{\text{true},j} \) are given in Equation A2. It follows from Equations A1 and A5 that the vector of observable scores also follows a bivariate normal distribution:

\[
\begin{bmatrix} Y \\ X \end{bmatrix} \mid P_j \sim N(\mu_j, \Sigma_j),
\]

where \( \mu_j \) is given in Equation A2, and \( \Sigma \) is given in Equation A4.

Using standard conditioning arguments, it can be shown that if the pair \((Y,X)\) is randomly drawn from population \( j \), then the distribution of \( Y \) conditional on \( X = x \) is the following:
\[ Y | (X = x, P_j) \sim N \left[ \beta_{0,j} + \beta_{1,j} x, \frac{\sigma_{y,j}^2}{\sigma_{x,j}} (1 - \rho_j^2 \alpha_{x,j} \alpha_{y,j}) \right]. \] (A6)

where

\[ \beta_{0,j} = \mu_{y,j} - \mu_{x,j} \beta_{1,j} \text{ and } \beta_{1,j} = \rho_j \alpha_{x,j} \frac{\sigma_{y,j}}{\sigma_{x,j}}. \]

**Case 2: Arbitrary Distribution for** \( X \)

In Case 2, the distributions of \( X_{true}, \varepsilon, \text{ and } X \) are arbitrary, provided that the moments in Equations A2 and A3 exist. No specific marginal distributions are assumed for either \( Y_{true} \) or \( Y \), but it is assumed that \( Y \) is conditionally normal given \( X = x \). Specifically, it is assumed that the regression model in Equation A6 holds.

**APPENDIX B**

Testing Hypotheses Regarding Moderating Effects

In testing moderating effect hypotheses a sample of \( n_j \) pairs is randomly selected from the \( j \)th population, for \( j = 1, \ldots, k \). These selections may be made completely at random, or they may be made according to a restricted random process. In any case, it is assumed that conditional on \( X = x \), the probability of selecting the pair \((Y, X)\) is independent of \( Y \). If this assumption is satisfied, then the conditional distribution of \( Y \) given \( X = x \) for sampled pairs \((Y, X)\) is identical to the conditional distribution of \( Y \) given \( X = x \) in the entire population (i.e., Eq. A6). One simple example of this type of restricted sampling is truncation on \( X \). That is, selection could be random subject to the restriction that each \((Y, X)\) pair in sample \( j \) satisfies \( X \geq x_j \) or \( X \leq x_j \), where \( x_j \) is a fixed lower or upper cutoff.

**Null Hypothesis**

To test the moderating effect hypotheses shown in Equation 1, one tests the hypothesis \( H_0: \beta_{11} = \beta_{12} = \ldots = \beta_{1k} \). This hypothesis is equivalent to

\[ H_0: \rho_1 \alpha_{x,1} \frac{\sigma_{y,1}}{\sigma_{x,1}} = \rho_2 \alpha_{x,2} \frac{\sigma_{y,2}}{\sigma_{x,2}} = \ldots = \rho_k \alpha_{x,k} \frac{\sigma_{y,k}}{\sigma_{x,k}}. \]

However, as reviewed by Aguinis and Pierce (1998a), a false null hypothesis could be attributable to heterogeneity among the correlations, the \( X \) variances, the \( Y \) variances, and/or the \( X \) reliabilities.

Denote the \( k \times 1 \) vector of slope parameters by \( \beta_j \). That is,
A vector of moderating effects can be obtained by computing contrasts among the entries of \( \beta \). Specifically, let \( C \) be a \( k \times (k-1) \) matrix of contrast coefficients with entries \( c_{ij} \) for \( i = 1, \ldots, k \) and \( j = 1, \ldots, k-1 \) defined as follows:

\[
c_{ij} = \begin{cases} 
1 & \text{if } i = j; \\
-1 & \text{if } i = k; \text{ and} \\
0 & \text{otherwise.}
\end{cases}
\]  

(B1)

For example, if \( k = 4 \), then

\[
C = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
-1 & -1 & -1
\end{pmatrix}
\]

Using this matrix notation, the vector of moderating effects is \( C' \beta \). For convenience, the vector slope differences, \( C' \beta \), will be denoted by \( \psi \). That is,

\[
C' \beta = \begin{pmatrix}
\beta_1 - \beta_k \\
\beta_2 - \beta_k \\
\vdots \\
\beta_{k-1} - \beta_k
\end{pmatrix} = \psi = \begin{pmatrix}
\psi_1 \\
\psi_2 \\
\vdots \\
\psi_{k-1}
\end{pmatrix}
\]

In short, the null hypothesis can be written either as \( H_0: C' \beta = 0 \) or as \( H_0: \psi = 0 \). (Note that one-degree-of-freedom contrasts also could be tested as opposed to an omnibus test given that a theory-based hypothesis exists regarding differences between/among specific groups; West, Aiken, & Krull, 1996.)

**Conventional MMR Model**

For convenience, the total sample size, \( \Sigma_{j=1}^k n_j \) is denoted by \( N \). In addition, the partial sum \( \Sigma_{j=1}^i n_j \) is denoted by \( N_i \) for \( j = 1, \ldots, k \); and \( N_i \) is defined as zero. That is, \( N_0 = 0; N_1 = n_1; N_2 = n_1 + n_2 \) and so forth up to \( N_k = N \). The pooled sample of pairs is \((Y, X)\) for \( i = 1, \ldots, N \), and the sample of size \( n \), from population \( j \), is \((Y, X)\) for \( i = 1 + N_1, 2 + N_2, \ldots, N_j \).

The moderator variable \( Z \) can take on the values 1, 2, \ldots, \( k \). The value of \( Z \) indicates the subpopulation from which a \((Y, X)\) pair was drawn. That is, if the pair \((Y, X)\) was drawn from subpopulation \( j \), then \( z = j \). For each \( z \), a set of \( k-1 \) binary indicator variables, \( w_{ij} \) for \( j = 1, \ldots, k-1 \), can be defined as follows:

\[
w_{ij} = \begin{cases} 
1 & \text{if } z = j; \\
0 & \text{otherwise.}
\end{cases}
\]  

(B2)
In the special case of a binary moderator variable (i.e., \( k = 2 \)), the subscript \( j \) on \( w_i \) can be omitted, and the binary indicator variable, \( w_i \), simplifies to

\[
\begin{align*}
\text{if } z_i = 1; \\
0 \text{ if } z_i = 2.
\end{align*}
\]

The conventional MMR model for \((Y_i, X_i), i = 1, \ldots, N\) can be written as follows:

\[
Y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{k-1} w_{ij} \tau_j + \sum_{j=1}^{k-1} w_{ij} \psi_j + \epsilon_i, \quad (B3)
\]

where \( w_i \) is defined in Equation B2, and the parameters of the MMR model are functions of the parameters of the conditional distributions in Equation A6. Specifically,

\[
\beta_0 = \beta_{0k}; \quad \beta_1 = \beta_{1k}; \quad \tau_j = \beta_{0j} - \beta_{0k} \text{ for } j = 1, \ldots, k - 1; \quad \text{and} \quad \psi_j = \beta_{1j} - \beta_{1k} \text{ for } j = 1, \ldots, k - 1.
\]

Also, \( \epsilon_i \) for \( i = 1, \ldots, N \) are independently distributed random errors. The random error terms are independently distributed as

\[
\epsilon_i | P_j \sim N \left[0, \frac{\sigma_i^2}{\alpha_{ij}} \right] \text{ for } i = 1 + N_j, 2 + N_j, \ldots, N_j.
\]

**F Test**

The conventional MMR F test for moderating variable effects is to reject \( H_0 \) if \( F_x \geq F_{k-1, N-2k, \alpha} \), where \( F_x \) is the computed value of the test statistic, and \( F_{k-1, N-2k, \alpha} \) is the critical value for a size \( \alpha \) test. The subscript \( x \) on \( F_x \) is a reminder that the test is conducted conditional on \( X = x \). The test statistic can be computed as follows:

\[
F_x = \frac{\hat{\Psi}^T C D_x C \hat{\Psi}}{(k-1)MSE}, \quad (B4)
\]

where \( C \) is defined in Equation B1; \( \hat{\Psi} \) is the ordinary least squares estimator of \( \Psi \) in the MMR model (Eq. B3);

\[
D_x = \text{Diag}(SSX_{j}^{-1}; j = 1, \ldots, k);
\]

\[
SSX_{j} = \sum_{i=1}^{N_j} (x_i - \bar{x}_j)^2; \quad \bar{x}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i;
\]

\[
MSE = \frac{\Sigma_{j} \Sigma_{i} SSE_j}{N - 2k} = \frac{\Sigma_{j} SSE_j}{N - 2k};
\]

\( \text{Diag}(\alpha; j = 1, \ldots, k) \) is a diagonal matrix given by
\[ \text{Diag}(a_j; j = 1, \ldots, k) = \begin{pmatrix} a_1 & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_k \end{pmatrix} \]

SSE is the sum of squared errors from the least squares fit of the MMR model (Eq. B3), and SSE\(_j\) for \(j = 1, \ldots, k\) is the sum of squared errors from the least squares fit of the parameters in the \(j\)th conditional model described in Equation A6.

### APPENDIX C

**Distribution of the Test Statistic Conditional on \(X\) and Approximation to the Unconditional Distribution of the Test Statistic**

#### Distribution of the Test Statistic Conditional on \(X\)

In general, the test statistic \(F_x\) will not be distributed as an \(F\) random variable. Under the model shown in Equation A6, the condition that must be satisfied for \(F_x\) to follow an \(F\) distribution is homogeneity of the conditional variances of \(Y\) given \(X\) (Aguinis & Pierce, 1998a). This homogeneity condition can be written as follows:

\[
\frac{\sigma_{y,j}^2}{\alpha_{y,j}}(1 - \rho_{y,j}^2 \alpha_{x,j} \alpha_{y,j}) = \sigma^2 \quad \text{for} \quad j = 1, \ldots, k, \tag{C1}
\]

where \(\sigma^2\) is a positive constant. If the homogeneity condition is not satisfied, then the \(F_x\) is not distributed as an \(F\) random variable. Nonetheless, the exact distribution of \(F_x\), conditional on \(X\), can be obtained by using well-known linear models theory (e.g., Stapleton, 1995). The result is summarized in Theorem 1.

#### Theorem 1: Conditional Distribution of \(F_x\)

Define \(V_x\) as

\[
V_x = \text{Diag} \left[ \frac{\sigma_{y,j}^2(1 - \rho_{y,j}^2 \alpha_{x,j} \alpha_{y,j})}{\alpha_{y,j} \text{SS}_{X,j}}, \quad j = 1, \ldots, k \right]
\]

and denote the eigenvalues and eigen-vectors of \((C'D_x)C'V_xC\) by \(\omega_{y,j}\) and \(u_{x,j}\) respectively. That is,

\[
(C'D_x)C'V_xCu_{x,j} = u_{x,j}\omega_{y,j}
\]

for \(j = 1, \ldots, k - 1\). The distribution of \(F_x\), conditional on \(X\), is
\[ F_x = \left( \frac{N - 2k}{k - 1} \right) \left( \sum_{j=1}^{k-1} \omega_{x,j} G_{x,j} \right) \left( \sum_{j=1}^{k} \frac{\sigma_{x,j}^2 (1 - \rho_{x}^2 \sigma_{x,j} \sigma_{x,j}^2) H_j}{\alpha_{x,j}} \right) \]

where \( G_{x,j} \) for \( j = 1, \ldots, k - 1 \) and \( H_j \) for \( j = 1, \ldots, k \) are independently distributed chi-squared random variables. Specifically, \( H_j \sim \chi^2(n_j - 2) \) for \( j = 1, \ldots, k \) and \( G_{x,j} \sim \chi^2(1, \lambda_{x,j}) \) for \( j = 1, \ldots, k - 1 \), where \( \lambda_{x,j} \) is a noncentrality parameter and is given by

\[ \lambda_{x,j} = \frac{(u'_{x,j} C \beta_j)^2}{2u'_{x,j} C V C u_{x,j}}. \]

In the special case of a binary moderator variable, the conditional distribution of \( F_x \) simplifies substantially. In this special case, \( k - 1 = 1 \);

\[ \omega_{x,1} = \frac{\sigma_{x,1}^2 (1 - \rho_{x}^2 \sigma_{x,1} \sigma_{x,1}^2)}{\alpha_{x,1} \sqrt{\text{SSX}_1}} + \frac{\sigma_{x,2}^2 (1 - \rho_{x}^2 \sigma_{x,2} \sigma_{x,2}^2)}{\alpha_{x,2} \sqrt{\text{SSX}_2}}, \]

and

\[ \lambda_{x,1} = \frac{(\beta_1 - \beta_2)^2}{2 \left( \frac{\sigma_{x,1}^2 (1 - \rho_{x}^2 \sigma_{x,1} \sigma_{x,1}^2)}{\alpha_{x,1} \sqrt{\text{SSX}_1}} + \frac{\sigma_{x,2}^2 (1 - \rho_{x}^2 \sigma_{x,2} \sigma_{x,2}^2)}{\alpha_{x,2} \sqrt{\text{SSX}_2}} \right)} \]

Conditional on the \( X \) variable, the power of the moderating effect tests is given by

\[ \text{Power}_x = \Pr(F_x \geq F_{k-1,N-k-1}). \quad (C2) \]

**Approximation to the Unconditional Distribution of the Test Statistic**

To obtain the unconditional power of the \( F \) test, the conditional power in Equation C2 must be averaged over all possible realizations of the \( X \) variable. That is, the unconditional power of the \( F \) test is

\[ \text{Power} = \mathbb{E}(\text{Power}_x), \quad (C3) \]

where \( \text{Power}_x \) is given in Equation C2, and the expectation is taken with respect to the distribution of \( X \). Exact analytic expressions for the unconditional power are not known. Accordingly, the unconditional power must be approximated.
The $X$ variable plays a role in the conditional power solely through $SSX_j$ for $j = 1, \ldots, k$. The strategy for the proposed approximation is straightforward. First, for each $j$, replace $1/SSX_j$ by its expectation. Second, use Theorem 1 and act as though $1/SSX_j = E(1/SSX_j)$ for $j = 1, \ldots, k$. This strategy is equivalent to expanding the unconditional power in Equation C3 in a Taylor series around $1/SSX_j = E(1/SSX_j)$ for $j = 1, \ldots, k$ and truncating the series after the linear term. With an error of order $n_j^{−2}$, the expectation of $1/SSX_j$ is

$$E\left(\frac{1}{SSX_j}\right) = \frac{n_j + 1}{(n_j - 1)E(SSX_j)}.$$

The relationship between $E(SSX)$ and $\sigma^2_{x,j}$ depends on the reliability of $X$ in subpopulation $j$ and on the manner in which $(Y, X)$ pairs were selected from the $j$th subpopulation. Let $Q$ be a random variable that takes on the value $Q = 1$ if the pair $(Y, X)$ can be selected from the population and takes on the value $Q = 0$ if the pair $(Y, X)$ cannot be selected from the population. If the $(Y, X)$ pairs are a simple random sample from the $j$th population, then $Q = 1$ for all pairs, and

$$E(S^2_{x,j}) = \frac{\sigma^2_{x,j}}{\alpha_{x,j}},$$

where $S^2_{x,j} = \frac{SSX_j}{n_j - 1}$ is the sample variance of $X$ in the $j$th sample. If the probability of selecting the pair $(Y, X)$ depends on $X$, then

$$E(S^2_{x,j}|Q = 1) = \frac{\sigma^2_{x,j}}{\alpha_{x,j}} \delta_j,$$

where $\delta_j = \frac{\text{Var}(X|P_j, Q = 1)}{\text{Var}(X|P_j)}$. (C4)

The quantity $\delta_j$ is a multiplying factor that depends on the distribution of $X$ and on the manner in which selection probabilities depend on $X$.

If sampling is restricted by left truncation on $X$, then

$$Q = \begin{cases} 
1 & \text{if } X \geq x^*, \text{ and} \\
0 & \text{otherwise.} 
\end{cases}$$

If the distribution of $X$ is known, then, for this simple form of restricted sampling, the multiplying factor $\delta_j$ can be computed explicitly. For example, if $X$ is normally distributed, then (using moment generating functions) the multiplying factor can be shown to be

$$\delta_j = 1 + \frac{\varphi(h_j)}{1 - T_j} \left( h_j - \frac{\varphi(h_j)}{1 - T_j} \right),$$

where $\varphi$ is the probability density function of the standard normal distribution, $h_j = \sqrt{\alpha_{x,j}} (x_j^* - \mu_j)/\sigma_{x,j}$; and $T_j$ is the truncation proportion for subpopulation $j$. That is, the sample consists of pairs $(Y, X)$ that are randomly selected from all pairs in population $j$ for which $X$ lies in the upper or lower $100[1 - T_j]$% of the distribution.

In the simple case of normally distributed $X$ and left truncation, the multiplying factor $\delta_j$ always is less than or equal to 1.00. In other cases, $\delta_j$ can be less than, equal to, or greater than 1.00. For example, suppose that the density function of $X$ in subpopulation $j$ is $f_j(x)$. Consider the selection mechanism
Prob(\(Q = 1|X = x\)) = \(F_{x}^{-1}(x)^{\frac{r}{T}}\), where \(F_{x}(x) = \int_{-\infty}^{x} f_{x}(u)du\)

is the cumulative distribution function of \(X\), and \(T\) is a constant in \((0, 1)\). With this selection mechanism, the probability of selection gradually increases as \(X\) increases. This selection mechanism is a continuous version of left truncation, and it will be called sparse left sampling. It can be shown that \(E[\text{Prob}(Q = 1|X)] = 1 - T\). That is, the percentage of the distribution that cannot be sampled is \(100T\%\). The value of \(T\), therefore, is analogous to the truncation proportion discussed above. The conditional density of \(X\) given \(Q = 1\) is

\[ f_{x|Q}(x|Q = 1) = \frac{1}{1 - T} \cdot f_{x}(x)F_{x}(x)^{-\frac{r}{T}}. \]

To compute \(\delta\), the density function \(f_{x}(x)\) must be known. As an example, suppose that \(X\) follows an exponential distribution in subpopulation \(j\). That is, the density function is

\[ f_{x|Q}(x|Q = 1) = \lambda e^{-\lambda x}, \text{ for } 0 < x < \infty. \]

It can be shown that \(\delta\) increases from \(\delta = 1\) to \(\delta = 1.645\) as the “truncation proportion” (i.e., \(T\)) increases from 0 to 1. That is, the sample variances tend to be inflated, and the degree of inflation increases as \(T\) increases.

In sparse right sampling, the probability of selection is high for low \(X\) scores, and it decreases as \(X\) increases. The probability of selection in sparse right sampling is

\[ \text{Prob}(Q = 1|X = x) = (1 - F_{x}(x))^{\frac{r}{T}}. \]

If an exponential distribution (in subpopulation \(j\)) is subjected to sparse right sampling, then \(\delta\) decreases from \(\delta = 1\) to \(\delta = 0.00\) as \(T\) increases from 0.00 to 1.00. If the selection mechanism is changed to the left truncation mechanism,

\[ \text{Prob}(Q = 1|X = x) = \begin{cases} 1 & \text{if } x > x^*, \text{ and} \\ 0 & \text{if } x < x^* \end{cases}, \]

then for the exponential distribution, \(\delta = 1\) regardless of the cutoff value \(x^*\). In summary, if \(X\) follows an exponential distribution, then \(\delta = 1, \delta < 1, \text{ or } \delta > 1\) depending on the sampling mechanism.

Next, Theorem 2 summarizes the power approximation obtained by making the required substitutions.

**Theorem 2: Unconditional Power Approximation**

The power of the MMR \(F\) test is

\[
\text{Power} = \Pr \left( \frac{k - 1}{N - 2k} F_{k - 1, N - 2k}^{1, -1} + \frac{\sigma_{x,j}^{2}}{\sum_{j=1}^{k} \alpha_{x,j}^{2}} \left( \frac{1 - \rho_{j}^{2}}{\alpha_{x,j}} \right) H_j \right) \leq \sum_{j=1}^{k} \omega_{j} G_{j} \leq 0 \right)
\]

where \(\omega_{j}\) is the \(j\)th eigenvalue of \((C^T D C)^{-1} C^T V C;\)
\(D = \text{Diag} \left[ \frac{\alpha_{a_j}(n_j + 1)}{(n_j - 1)^2 \delta \sigma^2_{x_{a_j}}}; j = 1, \ldots, k \right]\)

\(V = \text{Diag} \left[ \frac{\sigma^2}{\alpha_{a_j}(1 - \rho^2 \alpha_{a_j})\lambda^2(n_j + 1)}{\alpha_{a_j}(n_j - 1)^2 \delta \sigma^2_{x_{a_j}}}; j = 1, \ldots, k \right]\)

and \(G_j\) for \(j = 1, \ldots, k - 1\) and \(H_j\) for \(j = 1, \ldots, k\) are independently distributed chi-squared random variables. Specifically, \(H_j \sim \chi^2(n_j - 2)\) for \(j = 1, \ldots, k\) and \(G_j \sim \chi^2(1, \lambda_j)\) for \(j = 1, \ldots, k - 1\), where \(\lambda_j\) is a noncentrality parameter;

\[\lambda_j = \frac{(\mathbf{u}_j' \mathbf{C}\mathbf{B}_j)^2}{2(\mathbf{u}_j' \mathbf{C}\mathbf{V}\mathbf{C}\mathbf{u}_j)}\]

and \(\mathbf{u}_j\) is the \(j\)th eigen-vector of \((\mathbf{C}'\mathbf{D}\mathbf{C})^{-1}\).

The accuracy of the approximation in Theorem 2 increases as each \(n_j\) increases. Also, for any fixed sample size, the approximation is most accurate if the standardized kurtosis coefficients of the \(X\) variable are near zero (i.e., normal kurtosis) in each subpopulation.

**Notes**

1. We acknowledge that there is also a contrary view. For instance, Luce (1995) stated that evidence of interactions is usually a signal of trouble. . . All too often, in my opinion, the interactions are treated as a finding and not as evidence of a lack of understanding of the combining rule for measures of the independent variables. (p. 21)

   However, as noted by Aguinis (in press), Aguinis and Pierce (1998c), Aguinis and Whitehead (1997), and others (e.g., Hall & Rosenthal, 1991), it is the theory-based interaction effects that are “at the very heart of the scientific enterprise” (Hall & Rosenthal, 1991, p. 447). Alternatively, unexpected and/or unanticipated interaction effects can be problematic. They might lead to a meaningful discovery or simply indicate that the conception of the research question and/or design is incorrect.

2. Increasing sample size across subgroups is related but not equivalent to reduction of variance in the categorical predictor \(Z\). An increase in sample size can result in a decrease or an increase in the variance of \(Z\). For example, if \(Z\) is a categorical moderator that takes on values of 1 and 2, the sample variance of \(Z\) is

\[S^2_Z = \frac{\sum (Z_i - \bar{Z})^2}{N-1} = \frac{Np(1-p)}{N-1},\]

where \(N = n_1\) (i.e., sample size in Subgroup 1) + \(n_2\) (i.e., sample size in Subgroup 2), and \(p = n_1/N\). As an illustration, if \(n_1 = 10\) and \(n_2 = 40\), then the sample variance is \(8/49 = .1636\). If \(n_1\) is increased to 40, then the sample variance is \(20/79 = .253\) (i.e., an increase). If \(n_1\) remains at 10 and \(n_2\) is increased to 50, then the sample variance is \(25/177 = .1412\) (i.e., a decrease).

3. Note, however, that the computer program MMRPOWER that implements the analytic approximation allows for the inclusion of up to 20 levels or moderator-based subgroups (i.e., \(k \leq\)
Tables showing results for conditions $4 \leq k \leq 20$ are available from the authors. The pattern of results was similar to those reported herein for conditions $2 \leq k \leq 3$.

References


Herman Aguinis (http://www.cudenver.edu/~haguinis) is an associate professor of management at the University of Colorado at Denver. He received a Ph.D. in industrial/organizational psychology from the University of Albany, State University of New York. His current research interests include personnel selection, social power and influence, estimation of interaction effects, meta-analysis, and research methods.
Robert J. Boik (http://www.math.montana.edu/~rjboik) is a professor of statistics at Montana State University. He received a Ph.D. in experimental psychology from Baylor University and a Ph.D. in statistics from Temple University. His current research interests include linear models, multivariate statistics, and Bayesian methods.

Charles A. Pierce (http://www.montana.edu/wwwpy) is an associate professor of industrial/organizational psychology at Montana State University. He received a Ph.D. in social psychology from the University at Albany, State University of New York. His current research interests include workplace romance, sexual harassment, estimation of interaction effects, meta-analysis, and research methods.
Testing Interaction Effects in LISREL: Examination and Illustration of Available Procedures

JOSE M. CORTINA
GILAD CHEN
George Mason University
WILLIAM P. DUNLAP
Tulane University

The concomitant proliferation of causal modeling and hypotheses of multiplicative effects has brought about a tremendous need for procedures that allow the testing of moderated structural equation models (MSEMs). The seminal work of Kenny and Judd and Hayduk has been drawn on by several authors in the past 10 years, thus producing procedures that allow for such tests. Yet, utilization of MSEMs in empirical research has been quite rare. The purposes of this article are twofold. First, the authors discuss general issues with respect to multivariate normality, indicators of latent products, the nature of latent products, and identification problems in MSEM. Second, they review and illustrate techniques that are available for the testing of interaction effects in structural equation models.

As the social sciences have developed, the complexity of hypothesized relationships has increased steadily (Cortina, 1993). Two of the more obvious indicators of this complexity are the increasing frequency of hypotheses involving multiplicative effects (e.g., linear interaction effects, nonlinear effects) and the popularity of structural equations modeling (SEM). In spite of the preponderance of both multiplicative effects and structural equations models, there is considerable confusion about the appropriate methods for combining the two. In other words, there is confusion with respect to the manner in which multiplicative effects should be incorporated into covariance structures models (Hayduk, 1987; Mathieu, Tannenbaum, & Salas, 1992; Ping, 1995).

Strangely, this confusion is not due to a lack of methodology. There are a variety of techniques available for testing structural equations models with multiplicative terms (moderated structural equations models [MSEMs]), each with its own strengths and weaknesses. Nevertheless, most of these techniques are unknown outside mathemati-
cal and quantitative circles. The primary purpose of the present article is to review these techniques, describe their advantages and disadvantages, and provide illustrations of their use.

A secondary purpose of the present article is to discuss more general issues associated with MSEM. These issues include violation of the multivariate normality assumption, choice of indicators of the latent product term, determination of the nature of the product term, and avoidance of identification problems. These more general issues are discussed first.

We wish to make clear that our purpose is pedagogical. The majority of the substantive contents of this article have appeared elsewhere (e.g., Jöreskog & Yang, 1996; Li, Harmer, Duncan, & Boles, 1998; Ping, 1996a; Rigdon, Schumacker, & Wohtke, 1998; Schumacker & Marcoulides, 1998). For example, previous researchers have compared the parameter recovery and error rate properties of many of the available procedures. Cortina (1993) has reported that multiplicative hypotheses and SEM techniques are increasingly common. In spite of these facts, our review of various journals (e.g., Personnel Psychology, Journal of Applied Psychology, Academy of Management Journal) suggests that the number of studies testing multiplicative effects with SEM is only slightly larger than it was in the wake of Kenny and Judd’s (1984) seminal article on the topic.

One likely explanation for this phenomenon is the lack of user-friendly descriptions of these procedures. Many descriptions fail to include the LISREL code associated with the procedure. Those that do include code, such as Jöreskog and Yang (1996), Li et al. (1998), and Jonsson (1998), omit certain of the available procedures, rely on unnecessarily complicated estimation algorithms such as WLSA, and do not necessarily use the most recent LISREL codes or language. Also, and perhaps most important, these articles fail to explain the more abstruse operations in these programs. Because the rationale for many of these operations is less than obvious, the typical reader may not be able to make the modifications necessary to transport them to other sets of data. It is our intention to assemble previously published material into a coherent whole, including detailed illustrations, so that others can implement the procedures that we describe. Before moving on to these descriptions and illustrations, we discuss four issues that are likely to come up when testing multiplicative models in SEM.

**General issues in MSEM**

**Multivariate Normality**

The most common approach to parameter estimation in SEM is maximum likelihood (ML), if for no other reason than because it is, in comparison to other estimators, simple (i.e., a simpler weight matrix), both conceptually and computationally. However, one of the assumptions of ML is that the variables in the model are distributed multivariate normal. Unfortunately, inclusion of a product term violates this assumption (Kenny & Judd, 1984).

This fact has led to the development of “distribution-free” estimation techniques that make no assumptions about the distributions of the variables involved. For example, Browne (1984) suggested the use of a weighted least squares approach based on augmented moment matrices (WLSA option in LISREL: Jöreskog & Sorbom, 1993a).
It is this approach on which the Jonsson (1998) chapter is based. The primary difficulty with distribution-free estimation techniques is that they require considerably larger sample sizes than does ML estimation (Jöreskog & Sorbom, 1993a). With sample sizes that are typical of organizational research (i.e., < 200), estimation techniques such as WLSA are impractical because rank deficiency of the weight matrices leads to high risk for problems such as nonconvergence and multiple solutions (Jöreskog & Yang, 1996). In addition, Hu, Bentler, and Kano (1992) found that the asymptotically distribution-free estimator produced tremendously inflated model fit chi-squared values for samples as large as 1,000.

There are other alternatives to ML, such as generalized least squares, least squares based on elliptical distributions, and the Satorra-Bentler method involving correction for violation of distributional assumptions (Satorra & Bentler, 1988). However, there is considerable evidence that ML is robust with respect to many types of violation of the multivariate normality assumption (Bollen, 1989; Chou, Bentler, & Satorra, 1991). In particular, Bollen (1989) pointed out that if the exogenous indicators (Xs) are unrelated to the ζ values (the latent errors in the equations), and if the ζ variables are multivariate normal, then the usual properties of the ML estimator hold. It should be noted that these two requirements are nothing more than the latent variables model analogs of common ordinary least squares regression assumptions. Indeed, Hu et al. (1992) found that the ML estimator results in goodness-of-fit values that are similar to those produced by the robust Satorra and Bentler (1988) estimator, even in the presence of nonnormal kurtosis as long as factors and errors are independent. It is only when factors and errors are not independent that ML breaks down.

Finally, recent work by Jaccard and Wan (1995) suggested that ML estimation is superior in terms of bias in parameter estimates, Type I error rates, and power. More recent work by DeShon and Schmitt (1999) provided additional evidence for these claims. This work notwithstanding, there have been suggestions that ML will produce incorrect standard errors when the multinormality assumption is violated (Bollen, 1989). The degree to which this is the case is unknown, but we would suggest the following. First, models must be carefully specified. This is, of course, always important, but it is especially so for MSEM because omission of relevant variables is one of the most likely causes that relationship between exogenous indicators (Xs) and ζ values. Second, statistics for univariate and multivariate normality should be examined. It is the multivariate normality statistics that are most important, but because there are no commonly accepted benchmarks for these statistics, the univariate values can help in deciding how problematic violation of multinormality is likely to be. If there is evidence of considerable nonnormality (e.g., significant multivariate and univariate departures from normality), then one should interpret any statistics that are reliant on standard errors with care (e.g., significance test statistics, confidence bands, etc.: Jöreskog & Yang, 1996). In particular, any parameter estimates that are barely or marginally statistically significant must be viewed with more than the usual amount of suspicion. Of particular interest are departures from mesokurtosis. Although there exists a precise mathematical definition of mesokurtosis, it suffices to say that a normal distribution is mesokurtic as opposed to leptokurtic (i.e., tall and skinny) or platykurtic (i.e., short and fat). For example, Hu et al. (1992) used kurtosis values of –1, 2, and 5 for their nonmesokurtic factors and found that ML acquitted itself nicely under these conditions provided that factors and error were independent. This would suggest that variables yielding kurtosis values such as these are unlikely to produce the sorts of prob-
lems with ML estimation about which devotees of alternative estimation algorithms warn us.

If one has access to EQS or to the most recent versions of LISREL, one can employ the "robust estimator" developed by Satorra and Bentler (1988) in an attempt to obtain more precise standard errors. On the other hand, wayward standard errors presumably would have made their presence felt in the Monte Carlo studies mentioned above. It is possible that real data would behave differently from the contrived data included in Jaccard and Wan (1995) and DeShon and Schmitt (1999), but there is no formal evidence of such a phenomenon at this time. Given the evidence that does exist, it seems reasonable to proceed with ML estimation unless there is extreme nonnormality in the data and compelling reasons to believe that either the exogenous indicators (Xs) are correlated with ζ values (the latent error in the equations) or the ζ variables are multivariate nonnormal. In such a case, it may not be feasible to test the MSEM in LISREL. Because of the vagaries of MSEM, more research specific to MSEM is needed.

**Choice of Indicators of the Latent Product Term**

There have been a variety of suggestions with respect to the variables that might be used as indicators of a latent product variable. All of these suggestions involve the use of functions of the indicators of the main effect variables as indicators of the latent product. For example, suppose that the model being tested involves the effects of two latent variables, X and Z, and their product XZ on the latent variable Y. Suppose further that X is indicated by x1 and x2, and Z is indicated by z1 and z2. All of the procedures described below use as indicators of XZ combinations of the xs and zs. One of the primary differences among the procedures described below is in how they combine main effect indicators to produce indicators of the latent product. For example, Kenny and Judd (1984) recommended the use of all possible pairwise products of the main effect indicators (e.g., x1z1, x1z2, x2z1, x2z2), whereas Jöreskog and Yang (1996) recommended only one latent product indicator (e.g., x1z1). The differences between these methods are described in detail below. For the time being, it is enough to note that all of the procedures reviewed herein use as indicators of the latent product functions of the main effect indicators.

**Determination of the Nature of the Latent Product**

It is often unclear whether a latent product should be exogenous or endogenous, that is, whether determinants of the latent product should be specified in the model (endogenous) or assumed to lie outside the model (exogenous). If the latent components of the latent product XZ (i.e., X and Z) are both exogenous, then the latent product can be treated as exogenous, and relationships among these three latent variables are easily estimated. The issue becomes more difficult if one of the components of the latent products is endogenous. Suppose X is exogenous but Z is endogenous to a third predictor, W. Then, one can make a case for the notion that the latent product should be endogenous to W. If the latent product is endogenous, then its covariance with other variables cannot be specified in the model unless all variables are specified as y vari-
ables. Also, the notion of a variable affecting a latent product makes little sense conceptually; that is, there would seldom be any conceptual reason for including such a link in a causal model.

The solution to this problem lies in the recognition of the fact that the product term in most moderator analyses is not a variable with any conceptual meaning. Instead, it is a tool for examining a particular pattern of relationships among other variables. In the above example, XZ is used to investigate the effects of X and Z on Y. However, the product of X and Z has no place in the theoretical justification for the interactive effect. Typically, one would hypothesize that the effect of X on Y varies across levels of Z. There is no hypothesis that the XZ product or the partialled product has an effect on Y. Consider theories in which interaction hypotheses play a role (e.g., aptitude by treatment interactions in the ATI literature, goal difficulty by goal commitment in goal setting theory). In most cases, the theory has to do with the “main effect constructs” and the ways that they combine to influence behavior. When we say that those constructs combine “multiplicatively,” we mean that the influence of one on the behavior in question depends on the level of the other and nothing more. The term multiplicative comes more from the way that we test this sort of relationship than from anything else. This is important because it allows us to put the nature of the product term into perspective. In cases in which the latent product is an analytical tool, we need not concern ourselves with its external determinants. The “latent” product is not a construct in the strict sense of the term. It is a variable that can suffer from measurement error, and this measurement error must be taken into account when evaluating the interactive effect of two or more variables, but it is not a psychological entity in and of itself. Given that the product is merely a vehicle, it can be modeled as an exogenous variable in any circumstances.

This is not to say that latent products cannot be constructs. Subjective expected utility theory and expectancy theory are two examples that contain products that are themselves constructs. If such is the case, then incorporation of the product into a SEM analysis could be more problematic. Fortunately, most products in the social/organizational sciences are largely utilitarian.

**Avoiding Identification Problems in MSEM**

Because of the interrelatedness of the latent predictors in MSEM analysis, it is quite possible for the degrees of freedom in the model to drop below +1. When degrees of freedom drop to 0, the model is said to be just identified, and when the degrees of freedom drop below 0, the model is said to be underidentified. There are a variety of problems associated with underidentification, and these problems have been discussed elsewhere (see Hayduk, 1987, for a detailed description). For MSEM analyses, one can be confronted with degrees of freedom problems as a result of including paths linking the latent product and its indicators to all of the variables to which they are likely related. For example, the indicators of XZ may load on X or Z, and the errors associated with these indicators (the Θδ for the indicators of XZ) may be correlated with the errors associated with the indicators of X and Z. If these paths are omitted, then the fit indices associated with the model will suffer. If all such paths are included, the result can be an underidentified model.
One strategy for minimizing these problems is to begin the analysis by centering all observed variables. In regression, the use of mean-centered data (i.e., variables whose raw values have been replaced by deviation scores) removes nonessential ill conditioning; that is, centering variables prior to formation of products minimizes the relationships between the variables and the products created from them (Marquardt, 1980; Marquardt & Snee, 1975). In the case of MSEM, centering prior to formation of products minimizes the relationships between the indicators of $XZ$ and the indicators of $X$ and $Z$. Furthermore, if the indicators of $X$ and $Z$ are distributed multivariate normal, then the relationship between those indicators and products constructed from them is expected to be zero (Dunlap & Kemery, 1987). This would suggest little relationship between the latent product and the latent main effect variables. If these relationships are small, then there is little need to estimate them, resulting in more degrees of freedom and fewer identification problems.

Of course, centering does not necessarily reduce these relationships to a point at which they need not be estimated. It is still quite possible for substantial relationships between products and their components to remain after centering. As Jöreskog and Yang (1996) pointed out, observed variable means (i.e., those values that are directly affected by centering) are functions of a variety of parameters in an MSEM model. Indeed, these authors recommended the estimation of intercepts in addition to weights for MSEM models. If the magnitude of these relationships between products and their components necessitates their specification in the model, then other solutions to identification problems, such as reducing the number of indicators of the latent product (Jöreskog & Yang, 1996; Ping, 1995), must be sought. Nevertheless, there is no empirical work at present suggesting that centering is harmful within the context of ML estimation (see Jöreskog & Yang, 1996, for an alternative opinion). For these reasons, it is recommended that variables be centered prior to the formation of products in MSEM analyses.

Now that these more general issues have been discussed, we turn our attention to the specific methods that are available for analyzing MSEMs. The subsequent sections unfold as follows. First, the seminal work of Kenny and Judd (1984) and Hayduk (1987) is described. Second, the procedures suggested in Jöreskog and Yang (1996), Ping (1995, 1996b), Mathieu et al. (1992), and Jaccard and Wan (1995) are described and evaluated. Third, the matrix language and SIMPLIS language (Jöreskog & Sorbom, 1993b) versions of the LISREL 8 code for these procedures are offered for data described below. The input variance/covariance matrices for these analyses are contained in Table 1. Selected output is offered in Table 2. The purpose of offering this output is to allow the readers to corroborate their own analyses of our data with ours. Table 2 also contains, for comparative purposes, output from a moderated multiple regression analysis of these same data using unit-weighted composites and a simple product. Full output from all of these analyses is available from the authors on request. Also, although we focus on interaction effects only, the procedures described below can also be applied to nonlinear effects.

**Available Procedures for MSEM**

As understanding of the original work of Kenny and Judd (1984) has spread, and as existing programs have been developed to accommodate the requirements of models
with multiplicative effects, more procedures for testing such models have been
offered. Recent research has shown that the methods reviewed below recover param-
eter values with reasonable success (Jaccard & Wan, 1995; Jöreskog & Yang, 1996;
Ping, 1995). Although there appear to be some relatively small differences across
these methods with respect to the accuracy with which they recover parameter values,
an equally important dimension on which they appear to vary is usability. Indeed, what
good is a method that performs beautifully the function for which it was intended but
that requires exorbitant sample sizes or is too complicated and cumbersome for most

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariance matrix for Jöreskog and Yang’s (1996) procedure</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>X1</td>
</tr>
<tr>
<td>X2</td>
</tr>
<tr>
<td>Z1</td>
</tr>
<tr>
<td>Z2</td>
</tr>
<tr>
<td>X1Z1</td>
</tr>
</tbody>
</table>

| Covariance matrix for Ping’s (1995) procedure | Y | X1 | X2 | Z1 | Z2 | XZ |
| Y | .602 |
| X1 | -.268 | 1.052 |
| X2 | -.350 | .959 | 1.703 |
| Z1 | -.396 | .398 | .544 | .890 |
| Z2 | -.260 | .212 | .353 | .352 | 1.017 |
| XZ | .819 | -.404 | -.492 | -.778 | -.217 | 14.631 |

| Covariance matrix for Mathieu, Tannenbaum, and Salas’s (1992) procedure | Y | ZX1 | ZX2 | ZX1ZX2 |
| Y | NA | .602 |
| ZX1 | .849 | -.407 | 1.001 |
| ZX2 | .671 | -.287 | .433 | .999 |
| ZX1ZX2 | .638 | .234 | -.173 | -.122 | 1.193 |

| Covariance Matrix for Jaccard and Wan’s (1995) and Ping’s (1996) procedures | Y | X1 | X2 | Z1 | Z2 | X1Z1 | X1Z2 | X2Z1 | X2Z2 |
| Y | .602 |
| X1 | -.268 | 1.052 |
| X2 | -.350 | .959 | 1.703 |
| Z1 | -.396 | .398 | .544 | .890 |
| Z2 | -.260 | .212 | .353 | .352 | 1.017 |
| X1Z1 | .240 | -.183 | -.157 | -.171 | -.149 | 1.139 |
| X1Z2 | .161 | -.037 | -.027 | -.149 | .030 | .549 | 1.133 |
| X2Z1 | .270 | -.157 | -.188 | -.314 | -.144 | 1.012 | .587 | 1.705 |
| X2Z2 | .149 | -.027 | -.119 | -.144 | .045 | .562 | .995 | .717 | 1.810 |

(text continues on p. 334)
Table 2
Unstandardized Lambda and Phi Values and Fit Statistics From the Various Moderated Structural Equation Models Analyses on Sexual Harassment

1. Jaccard and Wan (1995) procedure: \( \chi^2 (df = 28) = 51.87; \text{RMSEA} = .05; \text{CFI} = .98; \text{AGFI} = .94 \)

<table>
<thead>
<tr>
<th>Construct</th>
<th>( C )</th>
<th>( S )</th>
<th>( C * S )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( Z_1 )</th>
<th>( Z_2 )</th>
<th>( X_1Z_1 )</th>
<th>( X_1Z_2 )</th>
<th>( X_2Z_1 )</th>
<th>( X_2Z_2 )</th>
<th>( Y ) (Sexual Harassment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate (C)</td>
<td>0.74</td>
<td></td>
<td></td>
<td>1.0*</td>
<td>1.27*</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.03</td>
</tr>
<tr>
<td>Supervision (S)</td>
<td>0.41*</td>
<td>0.57</td>
<td></td>
<td>—</td>
<td>—</td>
<td>1.0*</td>
<td>0.68*</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>−0.67*</td>
</tr>
<tr>
<td>( C * S )</td>
<td>—</td>
<td>—</td>
<td>0.59</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.0*</td>
<td>0.68*</td>
<td>1.27*</td>
<td>0.87*</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Theta Delta:</td>
<td>0.31</td>
<td>0.44</td>
<td>0.32</td>
<td>0.75</td>
<td>0.51</td>
<td>0.86</td>
<td>0.78</td>
<td>1.35</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Jöreskog and Yang (1996) procedure: \( \chi^2 (df = 9) = 8.71; \text{RMSEA} = .00; \text{CFI} = 1.0; \text{AGFI} = .99 \)

<table>
<thead>
<tr>
<th>Construct</th>
<th>( C )</th>
<th>( S )</th>
<th>( C * S )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( Z_1 )</th>
<th>( Z_2 )</th>
<th>( X_1Z_1 )</th>
<th>( Y ) (Sexual Harassment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate (C)</td>
<td>0.70</td>
<td></td>
<td></td>
<td>1.0*</td>
<td>1.38*</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.04</td>
</tr>
<tr>
<td>Supervision (S)</td>
<td>0.40*</td>
<td>0.57</td>
<td></td>
<td>—</td>
<td>—</td>
<td>1.0*</td>
<td>1.62*</td>
<td>−0.08*</td>
<td>−0.72*</td>
</tr>
<tr>
<td>( C * S )</td>
<td>—</td>
<td>—</td>
<td>0.56</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1.0*</td>
<td>0.27*</td>
<td></td>
</tr>
</tbody>
</table>
| Theta Delta:    | 0.36   | 0.38   | 0.33         | 0.80     | 0.55     | 0.28     |           |                         |}

(continued)
### Table 2 continued

3a. Ping (1995) procedure (two steps): \( \chi^2 (df = 7) = 16.50; \text{RMSEA} = .07; \text{CFI} = .98; \text{AGFI} = .95 \)

<table>
<thead>
<tr>
<th>Construct</th>
<th>C</th>
<th>S</th>
<th>C * S</th>
<th>Unstandardized Phi Coefficients</th>
<th>Lambda-X Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X1</td>
<td>X2</td>
</tr>
<tr>
<td>1. Climate (C)</td>
<td>1.0</td>
<td></td>
<td></td>
<td>0.83*</td>
<td>1.15*</td>
</tr>
<tr>
<td>2. Supervision (S)</td>
<td>0.64*</td>
<td>1.0</td>
<td></td>
<td></td>
<td>0.73*</td>
</tr>
<tr>
<td>3. C * S</td>
<td></td>
<td></td>
<td>1.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Theta Delta: 0.36</td>
<td>0.38</td>
</tr>
</tbody>
</table>

3b. Ping (1995) procedure (one step): \( \chi^2 (df = 8) = 16.64; \text{RMSEA} = .06; \text{CFI} = .98; \text{AGFI} = .95 \)

<table>
<thead>
<tr>
<th>Construct</th>
<th>C</th>
<th>S</th>
<th>C * S</th>
<th>Unstandardized Phi Coefficients</th>
<th>Lambda-X Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X1</td>
<td>X2</td>
</tr>
<tr>
<td>Climate (C)</td>
<td>1.0</td>
<td></td>
<td></td>
<td>0.84*</td>
<td>1.16*</td>
</tr>
<tr>
<td>Supervision (S)</td>
<td>0.64*</td>
<td>1.0</td>
<td></td>
<td></td>
<td>0.75*</td>
</tr>
<tr>
<td>C * S</td>
<td></td>
<td></td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Theta Delta: 0.36</td>
<td>0.38</td>
</tr>
</tbody>
</table>

4. Mathieu, Tannenbaum, and Salas (1992) procedure: \( \chi^2 (df = 2) = 8.29; \text{RMSEA} = .10; \text{CFI} = .97; \text{AGFI} = .93 \)

<table>
<thead>
<tr>
<th>Construct</th>
<th>C</th>
<th>S</th>
<th>C * S</th>
<th>Unstandardized Phi Coefficients</th>
<th>Lambda-X Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ZX1</td>
<td>ZX2</td>
</tr>
<tr>
<td>Climate (C)</td>
<td>1.0</td>
<td></td>
<td></td>
<td>.82a</td>
<td></td>
</tr>
<tr>
<td>Supervision (S)</td>
<td>0.57</td>
<td>1.0</td>
<td></td>
<td>0.92a</td>
<td></td>
</tr>
<tr>
<td>C * S</td>
<td></td>
<td></td>
<td>1.19</td>
<td></td>
<td>0.80a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Theta Delta: 0.33</td>
<td>0.15</td>
</tr>
</tbody>
</table>
5. Ping (1996) procedure: $\chi^2 (df = 28) = 57.02; \text{RMSEA} = .05; \text{CFI} = .97; \text{AGFI} = .94$

<table>
<thead>
<tr>
<th>Construct</th>
<th>C</th>
<th>S</th>
<th>C * S</th>
<th>X1</th>
<th>X2</th>
<th>Z1</th>
<th>Z2</th>
<th>X1Z1</th>
<th>X1Z2</th>
<th>X2Z1</th>
<th>X2Z2</th>
<th>Y (Sexual Harassment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate (C)</td>
<td>0.65</td>
<td></td>
<td></td>
<td>1.0*</td>
<td></td>
<td>1.36*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.001</td>
</tr>
<tr>
<td>Supervision (S)</td>
<td>0.40*</td>
<td>0.65</td>
<td></td>
<td></td>
<td>1.0*</td>
<td>0.67*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.61*</td>
</tr>
<tr>
<td>C * S</td>
<td></td>
<td>0.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theta Delta:</td>
<td>0.35</td>
<td>0.41</td>
<td>0.30</td>
<td>0.78</td>
<td>0.56</td>
<td>0.91</td>
<td>0.78</td>
<td>1.42</td>
<td></td>
<td></td>
<td></td>
<td>0.32</td>
</tr>
</tbody>
</table>

Moderated multiple regression analysis of sexual harassment data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Unstandardized Regression Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.Climate composite (C)</td>
<td>-.327*</td>
</tr>
<tr>
<td>2.Supervision (S)</td>
<td>-.129*</td>
</tr>
<tr>
<td>3.C * S</td>
<td>.135*</td>
</tr>
<tr>
<td>Constant</td>
<td>.506</td>
</tr>
</tbody>
</table>

Note. $N = 300$. RMSEA = root mean square error of approximation; CFI = comparative fit index; AGFI = adjusted goodness-of-fit index. Weights for latent products are in bold and underlined.

a. Indicator was used to set a factor value to unity and/or was constrained.

* $p < .05.$
researchers to actually use? This question is particularly relevant for the testing of multiplicative effects in structural equation models. For example, many researchers have felt a need for methods that allow tests of multiplicative effects for some time, and the Kenny and Judd method has been available for much of this time. Yet, we are unaware of a single instance of its implementation! There appear to be no mathematical or conceptual problems with the Kenny and Judd procedure excepting perhaps the lack of estimation of intercepts (cf. Jöreskog & Yang, 1996), and it certainly was not presented in an obscure journal. Nevertheless, the procedure in its original form was not used, and the reasons appear to be that it was too complicated, required too deep an understanding of SEM and its assumptions, placed unreasonable demands on the design of the experiment (e.g., prohibitive sample size) to be useful to any but an elite few, and failed to provide necessary statistics such as standard errors.

At the opposite pole are procedures such as multiple-groups analysis (MGA). MGA can be used when the moderator variable in question is categorical. In this case, a separate model is run for each level of the categorical moderator, and the models are compared with respect to path coefficients and model fit. Although this procedure is implemented with relative ease, it has three serious limitations. First, it requires additional external computations to generate path coefficients from product terms to the endogenous variable of interest and the significance tests associated with those coefficients. Thus, the values of primary interest cannot be estimated directly. Second, because it allows for no continuous moderator variables, many moderators must first be artificially categorized, resulting in loss of information and nonlinear, nonrandom measurement error. Third, separate analyses within groups require an assumption of perfect measurement in the categorical variable because its error cannot be modeled in the same way that the error associated with other variables can be modeled. In spite of these limitations, this procedure is relatively widely used (see Rigdon et al., 1998, for examples of implementation of this procedure).

It is certainly critical that a given procedure produce reasonably efficient, unbiased parameter estimates, but efficiency and lack of bias are insufficient for the use of complicated procedures to become common. Much of the difficulty associated with the testing of multiplicative effects in structural equation models derives from the fact that many researchers are unable to translate the typical descriptions of the available procedures into a usable set of command lines in programs like LISREL. For this reason, our descriptions of the available procedures include not only the conceptual foundations of the procedures but also actual command lines that would be used to analyze a particular model in LISREL 8 using both the traditional matrix language and the relatively new SIMPLIS language.

**Theoretical Foundations: Kenny and Judd (1984)**

No treatment of MSEM would be complete without some discussion of the work of Kenny and Judd (1984). These authors provided the foundation for all of the procedures developed since. To illustrate their approach, consider the model presented in Figure 1. In this model, the effect of organizational climate toward sexual harassment on experiences of sexual harassment depends on supervisory support. To test this hypothesis, we must, in some way, create a fourth variable that is the product of the climate and supervisory support variables. Thus, the model that we will test might look something like the model in Figure 2.
Here, we have three exogenous variables, climate ($\xi_1$), supervisory support ($\xi_2$), and Climate $\times$ Supervisory Support (the product of $\xi_1$ and $\xi_2$), which we will call $\xi_3$, affecting a fourth variable, sexual harassment experiences ($\eta$). Also, it is likely that climate and supervisory support will covary with one another and with their product. The
values representing the relationships among the exogenous concepts would be labeled \( \phi_{12} \), \( \phi_{13} \), and \( \phi_{23} \). Suppose that we have two indicators of climate, \( X_1 \) and \( X_2 \), and two indicators of supervisory support, \( X_3 \) and \( X_4 \). Kenny and Judd (1984) suggested the use of all possible cross products of the existing indicators as indicators of the latent product. In the above model, we would have four indicators of the latent product, \( X_1 \times X_3 \) (which we will call \( X_5 \)), \( X_2 \times X_3 \) (which we will call \( X_6 \)), \( X_1 \times X_4 \) (which we will call \( X_7 \)), and \( X_2 \times X_4 \) (which we will call \( X_8 \)). Thus, we have eight indicator variables, plus a ninth \( (Y_i) \) that will serve as the sole indicator of the endogenous variable \( \eta \). We would also need to incorporate the measurement error variances and covariances associated with the model \( (\theta) \), but these are left out of the figure to avoid confusion.

At first glance, this may seem like a fairly simple model. As Kenny and Judd (1984) pointed out, however, this is not the case. The difficulty lies in the computations associated with the indicators of the latent product. These indicators are, of course, composites of the indicators of climate and supervisory support. As such, the indicators of the latent product are functions of climate, supervisory support, and the errors associated with their indicators. Hayduk (1987) explained this issue by expressing the indicators of the simple latent variables (climate and supervisory support) as follows:

\[
X_1 = 1.0 \times \xi_1 + \epsilon_1
\]

\[
X_2 = \lambda_{21} \times \xi_1 + \epsilon_2
\]

\[
X_3 = 1.0 \times \xi_2 + \epsilon_3
\]

\[
X_4 = \lambda_{42} \times \xi_2 + \epsilon_4
\]

where the \( \lambda \)s are the path coefficients representing the relationships between exogenous constructs and their indicators, and the \( \epsilon \)s are errors. The \( \lambda \)s for \( X_1 \) and \( X_2 \) are often set to unity to define the scale. That is, one of the ways of coping with the problem that latent variables have no inherent measurement scale is to arbitrarily assign them one. By fixing the path from a latent variable to one of its indicators equal to 1, we automatically give that latent variable the same scale as that indicator. If no such specification is made, LISREL assigns a variance of 1 to latent variables by default.

\( X_5 \) through \( X_8 \) can then be represented as functions of Equations 1 through 4:

\[
X_5 = (1.0 \times \xi_1 + \epsilon_1) \times (1.0 \times \xi_2 + \epsilon_3) = \xi_1 \xi_2 + \xi_1 \epsilon_3 + \xi_2 \epsilon_1 + \epsilon_1 \epsilon_3
\]

\[
X_6 = (1.0 \times \xi_1 + \epsilon_1) \times (\lambda_{42} \times \xi_2 + \epsilon_4) = \lambda_{42} \xi_1 \xi_2 + \lambda_{42} \xi_1 \epsilon_4 + \lambda_{42} \xi_2 \epsilon_1 + \epsilon_1 \epsilon_4
\]

\[
X_7 = (\lambda_{21} \times \xi_1 + \epsilon_2) \times (1.0 \times \xi_2 + \epsilon_3) = \lambda_{21} \xi_1 \xi_2 + \lambda_{21} \xi_1 \epsilon_3 + \lambda_{21} \xi_2 \epsilon_1 + \epsilon_2 \epsilon_3
\]

\[
X_8 = (\lambda_{21} \times \xi_1 + \epsilon_2) \times (\lambda_{42} \times \xi_2 + \epsilon_4) = \lambda_{21} \lambda_{42} \xi_1 \xi_2 + \lambda_{21} \lambda_{42} \xi_1 \epsilon_4 + \lambda_{21} \lambda_{42} \xi_2 \epsilon_1 + \epsilon_2 \epsilon_4
\]
As can be seen, these indicators of the latent product are complex functions of the two latent variables, their links to their indicators, and the errors associated with those indicators. The difficulty lies primarily in estimating the variances and covariances associated with complex terms such as $\lambda, \xi, \varepsilon$, and $\epsilon$. The equations for these complex variances include nonlinear terms such as squared loadings (e.g., $\lambda^2$). Although a complete description of the process involved in estimating these values is beyond the scope of this article (see Hayduk, 1987; Kenny & Judd, 1984), it suffices to say that, at the time that Kenny and Judd (1984) and Hayduk (1987) were written, there was no way to include nonlinear constraints in programs such as LISREL. Because the solutions to this problem were complicated, it is not surprising that the approaches outlined in Kenny and Judd and Hayduk have been seldom applied to real data.

Much of the difficulty in the Kenny and Judd (1984) and Hayduk (1987) approaches is created to some extent by the fact that all possible cross products of the indicators of the latent variables are used as indicators of the latent product. As an alternative, some authors have suggested the use of single indicators of the latent product (e.g., Jöreskog & Yang, 1996; Mathieu et al., 1992; Ping, 1995), whereas others have suggested a reduced number of indicators (Jaccard & Wan, 1995). Furthermore, these methods are shown to require a similar set of assumptions as does the Kenny and Judd approach, and they have generally been shown to recover known parameter values with reasonable success (Jaccard & Wan, 1995; Jöreskog & Yang, 1996; Ping, 1995). We now discuss five such approaches. These approaches are demonstrated using data from a large-scale study of predictors of sexual harassment experiences.

Empirical Demonstrations

The Data Set and the General Approach

The purpose of the study that produced the data that we use to demonstrate the available MSEM procedures was to predict sexual harassment experiences from additive and multiplicative combinations of climate for sexual harassment and supervisory support. Larger values for climate indicate tolerance of sexual harassment, and larger values for supervision indicate a supervisor who is supportive of harassed or potentially harassed subordinates. Input variance/covariance matrices from the sample of 300 participants are provided in Table 1. The mean age in the sample is 31.2 years ($SD = 7.14$), 78% were women, and 67% were White. Because each approach requires different input, a different matrix is offered for each. To keep matters simple, all-$X$ models are used to demonstrate the different procedures. That is, all latent concepts in the examples that follow are treated as exogenous ($\xi$). This is accomplished by using a single indicator of the dependent variable and treating the loadings of this variable onto the exogenous concepts as path coefficients. Thus, the values carrying the effects of the exogenous variables on the dependent variable, usually represented as $\gamma$ values, are actually $\lambda$ values in the model tested here. The use of all-$X$ models simplifies the illustration by eliminating the need for $\lambda$, $\beta$, $\theta$, $\gamma$, and $\psi$ matrices, thus allowing attention to be focused on the various procedures and the differences among them (see Jöreskog & Yang, 1996, or Ping, 1996b, for other examples). One need simply interpret $\phi$ and $\theta$, and most $\lambda$ values normally, and interpret $\lambda$ values relating latent variables to $y$ as if they were gamma values. In any case, the use of endogenous latent vari-
ables and their measurement models would not affect the implementation of these procedures.

In the sections that follow, descriptions of the Jaccard and Wan (1995), Jöreskog and Yang (1996), Ping (1995), Mathieu et al. (1992), and Ping (1996b) procedures are offered as well as demonstrations using the aforementioned data set. The similarities and differences in the output associated with these different procedures are also discussed.

Before moving on to these descriptions, it may be worthwhile to remind the reader that some values are fixed by default in LISREL, whereas others are freely estimated by default. In particular, all phi values (variances and covariances for latent exogenous variables) are freely estimated, as are all diagonal values in the theta delta matrix (error variances for the indicators of latent exogenous variables). Any such values that are to be fixed at zero, fixed at some nonzero value, or constrained to equal other values from the output must be dealt with explicitly (e.g., lines 9-10 in Table 3). All lambda X values (loadings of indicators onto exogenous latent variables) and off-diagonal values in theta delta (error covariances) are set to zero by default. Any such values that are to be estimated or fixed to a nonzero value must be dealt with explicitly (e.g., line 6 in Table 3).

**Jaccard and Wan (1995)**

The first approach to be described here has much to recommend it. The approach outlined in Jaccard and Wan (1995) is essentially the Kenny and Judd (1984) approach using the latest version of LISREL. LISREL 8 allows the specification of the nonlinear constraints required by MSEM models. For this reason, one need not be concerned with convenience variables (cf. Hayduk, 1987). Also, there is less of a need to minimize the number of indicators of the latent product, although it should be noted that a large number of indicators relative to sample size can result in an unstable observed covariance matrix. In any case, the logic of the Jaccard and Wan procedure is identical to that of the Kenny and Judd procedure but simpler to implement. The model that is tested is identical to the model presented in Figure 2 with three exceptions. First, the path from each latent variable to its first indicator is fixed at 1 to define the scale of that latent variable. Second, because this is an all-X model, there would be arrows from the ksis (ξi) to the Y variable instead of to an eta (ηi). Third, the gammas (γs) would be replaced by lambdas (λs). ξ1 would represent climate for harassment, ξ2 would represent supervisory support, ξ3 would represent the latent product, and Y would be the single indicator of sexual harassment experiences. This model is presented in Figure 3.

The command lines for analysis of the previously described data set are presented in Table 3. Recall that the primary hypothesis being tested is that the climate and supervisor variables interact to affect reports of sexual harassment experiences. The notes to Table 3 provide explanations for the more abstruse operations in the program. In particular, it should be noted that (a) the errors for the indicators of the latent product are allowed to correlate because these indicators share components (line 6); (b) the relationships between the latent product and its latent components are fixed to zero because centering should minimize these relationships (lines 9-10); (c) as is common in structural equation models, the loading of one indicator per latent variable is fixed to a value of 1 so that the latent variable in question takes on the scale of that indicator (lines 11-12); and (d) the variance of the latent product, the loadings of the indicators of the latent product onto the latent product, and the error variances for those indicators...
are constrained to equal values suggested by equations from various sources (lines 13, 14-16, and 17-20, respectively).

Unstandardized coefficients and fit indices for this analysis are contained in Table 2. The reason for preferring unstandardized path coefficients has to do with the fact that the variance of the latent product \((PH_{3,3})\) is a function of the variances of the latent main effect variables and their covariance. As such, this variance is bound to be different from the variances of the latent main effect variables. Unfortunately, the variance of the latent product in the standardized LISREL solution is always 1 and, therefore, incorrect. Only the unstandardized value is correct.\(^4\)

It should first be noted that, although the chi-squared value for this model is statistically significant, this is not cause for concern given the sample size of the analysis. The other fit statistics reported are well within the bounds normally considered to be acceptable.

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) DA NI=9 NO=300</td>
</tr>
<tr>
<td>2) LA</td>
</tr>
<tr>
<td>3) Y X1 X2 Z1 Z2 X1Z1 X1Z2 X2Z1 X2Z2</td>
</tr>
<tr>
<td>4) CM FI=KJ2.CM</td>
</tr>
<tr>
<td>5) MO NX=9 NK=3 TD=SY PH=SY LX=SY</td>
</tr>
<tr>
<td>6) FR TD(6,7) TD(6,8) TD(7,9) TD(8,9)(^a)</td>
</tr>
<tr>
<td>7) FR LX(1,1) LX(1,2) LX(1,3) LX(3,1) LX(5,2) LX(7,3) LX(8,3)</td>
</tr>
<tr>
<td>8) FR PH(1,1) PH(2,2) PH(3,3) PH(2,1)</td>
</tr>
<tr>
<td>9) FI PH(3,1) PH(3,2)(^b)</td>
</tr>
<tr>
<td>10) VA 0 PH(3,1) PH(3,2)(^b)</td>
</tr>
<tr>
<td>11) FI LX(2,1) LX(4,2) LX(6,3)(^c)</td>
</tr>
<tr>
<td>12) VA 1 LX(2,1) LX(4,2) LX(6,3)(^c)</td>
</tr>
<tr>
<td>13) CO PH(3,3)=PH(1,1)*PH(2,2)+PH(2,1)**2(^d)</td>
</tr>
<tr>
<td>14) EQ LX(7,3)=LX(5,2)(^e)</td>
</tr>
<tr>
<td>15) EQ LX(8,3)=LX(3,1)</td>
</tr>
<tr>
<td>16) CO LX(9,3)=LX(3,1)*LX(5,2)</td>
</tr>
<tr>
<td>17) CO TD(6,6)=PH(1,1)**TD(4,4)+PH(2,2)**TD(2,2) +TD(4,4)**TD(2,2)(^f)</td>
</tr>
<tr>
<td>18) CO TD(7,7)=PH(1,1)**TD(5,5) +LX(5,2)**2*PH(2,2)**TD(2,2)+TD(5,5)**TD(2,2)</td>
</tr>
<tr>
<td>19) CO TD(8,8)=LX(3,1)**2*PH(1,1)**TD(4,4)+PH(2,2)**TD(3,3)+TD(4,4)**TD(3,3)</td>
</tr>
<tr>
<td>20) CO TD(9,9)=LX(3,1)**2<em>PH(1,1)**TD(5,5)+LX(5,2)**2</em>PH(2,2)**TD(3,3)+ C</td>
</tr>
<tr>
<td>20) PD</td>
</tr>
<tr>
<td>21) OU AD=OFF IT=100</td>
</tr>
</tbody>
</table>

a. Errors for the product indicators should correlate with each other because the indicators share components.  
b. Lines 9 and 10 fix the relationships between the latent product and its latent components at zero because they should be near zero as a result of centering.  
c. These values are fixed at 1 to define the scales of the latent variables.  
d. This serves to set the variance of the latent product equal to the product of the variances of its components plus the square of their covariance, as per Hayduk (1987, Eq. 7.44).  
e. The operations in lines 14 through 16 represent the constraining of the paths from the latent product to its indicators \(\lambda_{ij}\) equal to \(\lambda_{ij}\)\(^2\), as suggested by Jaccard and Wan. However, in the case of \(LX(7,3)\) and \(LX(8,3)\), one of the relevant \(\lambda\) values was fixed at 1. Thus, each of these two values is simply equal to the relevant \(\lambda\) value that was not fixed at 1.  
f. Lines 17 through 20 constrain the variances of the indicators of the latent product as per Equations 7 through 13 in Jaccard and Wan.
The only off-diagonal value in the phi matrix that is estimated is the relationship between the two latent main effect variables. This value, .41, shows that there is a statistically significant relationship between the climate and supervision latent variables.

Relevant lambda-X values are also contained in the table. Some of these values were fixed to define the scales of latent variables (latent variables have no inherent scale). All of those values that we estimated are significantly different from zero, thus providing support for the measurement model. Of particular interest are the values in the last column of the Jaccard and Wan portion of Table 2. Here, we have the loadings of the sexual harassment variable onto the three latent variables as well as its error variance (theta delta) value. Not surprisingly, there is a significant negative loading for the supervision variable. In other words, the more supportive the supervisor is, the less likely it is that the respondent experienced harassment. The loading for the climate variable is nonsignificant. However, there is also a significant loading for the latent product. This suggests that the effect of climate on likelihood of harassment depends on the degree of supervisor support. Specifically, the positive loading (.21) suggests that a supportive supervisor can mitigate the harmful effects of a hostile climate.

To interpret this interaction further, a plot of the interaction is presented in Figure 4. This plot was created by adapting the procedure described in Aiken and West (1991) using the standardized path coefficients. We used the standardized equation because the intercept for the unstandardized equation can only be generated from the use of mean structures, which are necessary only in the Jöreskog and Yang (1996) procedure. It seems unlikely that many researchers will go to the trouble of implementing the
Jöreskog and Yang procedure merely to generate an intercept value that locates lines of best fit on the Y-axis. Thus, our plots reflect the standardized equations. Nevertheless, the ideal technical approach would be to base all interpretation on unstandardized coefficients.

As can be seen, this plot suggests that a hostile climate results in more negative harassment experiences in the absence of supervisor support but has little effect on harassment experiences in the presence of strong supervisor support.

Jaccard and Wan (1995) showed that their procedure, with ML estimation, recovers parameter values better than does the corresponding distribution-free procedure and yields lower Type I error rates and higher power. Thus, the Jaccard and Wan approach is likely to be at the high end of the parameter recovery dimension. Because the logic of the procedure is the same as that of Kenny and Judd (1984), it is no more difficult (or easy) to understand. Nevertheless, the complexity associated with the use of complex products and multiple indicators of the latent product is likely to create convergence problems for some data sets. Also, the Jaccard and Wan procedure involves complex products that cannot be incorporated into the SIMPLIS language at this time. Thus, although the approach is certainly more usable than is the Kenny and Judd approach, it may be less tractable than the Ping (1995) or Mathieu et al. (1992) procedures.

**Jöreskog and Yang (1996)**

Jöreskog and Yang (1996) suggested that a single cross-product indicator be used for the latent product. This simplifies the analysis considerably. However, these authors suggested that mean structures should then be included in the analysis (see Jöreskog and Sorbom, 1993a, chap. 10, and see Ping, 1998, for an alternative procedure). For the typical analysis of an additive model, mean structures are not necessary.
if one centers observed variables prior to analysis, thus reducing intercept terms to zero. When multiplicative effects are involved, some of the observed variables are functions of other variables. As a result, centering fails to reduce all intercepts to zero (Jöreskog & Yang, 1996). Mean structures carrying intercept information can offset this difficulty. This implies that the structural equations analysis is defined by slightly different equations. Specifically, we add intercept terms $\alpha$, $\tau_x$, and $\tau_y$ to the traditional equations, such that

$$\eta = \alpha + \beta^*\eta + \Gamma\xi + \zeta,$$  \hspace{1cm} (9)

$$y = \tau_y + \Lambda_y\eta + \epsilon,$$  \hspace{1cm} (10)

$$x = \tau_x + \Lambda_x\xi + \delta.$$  \hspace{1cm} (11)

It is these intercept terms that allow one to estimate the values associated with the products in the model. The model that is actually tested with the Jöreskog and Yang (1996) approach is presented in Figure 5.

In addition to the hypothesized paths from $\xi_1$, $\xi_2$, and $\xi_3$ to their respective indicators and to the dependent variable, $y$, there are also paths from $\xi_1$ and $\xi_2$ to the indicator of the latent product. The coefficients for these paths are set equal to $\tau_1$ and $\tau_2$, which are the intercept terms associated with the indicators whose product makes up the single indicator of the latent product. This allows for the estimation of the complex variances that are present in such models without the creation of convenience variables.

As was discussed earlier, ML estimation is preferable for a variety of reasons. It does not call for a complicated weight matrix (e.g., the asymptotic parameter estimate variance/covariance matrix), and its sample size requirements are less severe. Instead, the ML version of the Jöreskog and Yang (1996) procedure requires a vector of sample means and specifications for the mean structures matrices $\tau_x$, $\tau_y$, and $\kappa$, where the $\tau$ values are as presented in Equations 10 and 11, and $\kappa$ contains the means of the exogenous concepts. To illustrate, matrix language command lines for a LISREL 8 run of the example described earlier are presented in Table 4. For the model in Figure 5, there are no endogenous concepts, so $\tau_y$ is not needed. Once again, explanations for the more difficult operations are contained in the notes to the table. Of particular note are (a) the use of a single indicator for the latent product (line 6); (b) the presence of the kappa vector, which contains latent variable means (lines 6, 10, and 17); (c) the presence of tau-X, which contains intercept values for the measurement model equations (line 6); and (d) the use of tau-X values to generate various values associated with the indicator of the latent product (lines 11-14, 16, and 18). Lines 7, 8, 9, and 15 carry the same functions as did the corresponding lines in the Jaccard and Wan (1995) run.

Table 2 contains relevant output from the analysis of the sexual harassment data using the Jöreskog and Yang (1996) procedure. As with the run based on the Jaccard and Wan (1995) procedure, the measurement model values from the Jöreskog and Yang procedure are encouraging, with all estimated values substantial and significantly different from zero. Also, the weight for the latent product is positive and significant. Thus, the Jöreskog and Yang analysis suggests once again that supervisory sup-
port mitigates the negative impact of a hostile climate. Figure 6 contains a plot of this interaction that is nearly identical to that produced with the Jaccard and Wan results and would suggest the same interpretational language. We are reluctant to make more specific comparative statements because of the fact that the comparison is based on a single data set.

Some differences between the Jöreskog and Yang (1996) and Jaccard and Wan (1995) approaches can also be seen. Specifically, the fit statistics for the Jöreskog and Yang procedure are better, and the weight for the latent product is larger. The difference in fit statistics should come as no surprise given the difference in overidentifying restrictions (28 vs. 9). Fit statistics such as the AGFI are intended to account for lack of parsimony, but the penalty for lack of parsimony associated with the AGFI is fairly small (James, Mulaik, & Brett, 1982). Although not reported here, fit statistics that penalize more heavily for lack of parsimony such as the PGFI and PNFI are higher for the analysis using Jaccard and Wan than for the analysis using Jöreskog and Yang.

As for the weight for the latent product, this difference is not large. Given the similarity of the plots of these interactions, it seems safe to say that, at least in the case of these data, these two procedures produce similar results.

Jöreskog and Yang (1996) show that this procedure recovers parameters efficiently and does not require nearly as many of the nonlinear constraints demanded by the Kenny and Judd (1984) approach. Also, the inclusion of mean structures adds further to the precision of the estimates. We would add, however, that this approach is somewhat unwieldy. The complexity created by the inclusion of mean structures makes
convergence more difficult, which in turn would mean that the approach might not yield parameter estimates for many data sets. This does not necessarily indicate that the procedure will go unused, but if the lack of application of the Kenny and Judd procedure is any indication of the likelihood that researchers will employ cumbersome techniques, it seems unlikely that the Jöreskog and Yang approach will find favor with many researchers. In addition, commands containing complex products, such as line

### Table 4


<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>DA NI=6 NO=300</td>
</tr>
<tr>
<td>2)</td>
<td>LA</td>
</tr>
<tr>
<td>3)</td>
<td>Y X1 X2 Z1 Z3 X1Z1</td>
</tr>
<tr>
<td>4)</td>
<td>ME=J&amp;Y96.ME</td>
</tr>
<tr>
<td>5)</td>
<td>CM=J&amp;Y96.CM</td>
</tr>
<tr>
<td>6)</td>
<td>MO NX=6 NK=3 TD=SY TX=FR KA=FR</td>
</tr>
<tr>
<td>7)</td>
<td>FR LX(1,1) LX(1,2) LX(1,3) LX(3,1) LX(5,2)</td>
</tr>
<tr>
<td>8)</td>
<td>FI PH(3,1) PH(3,2)</td>
</tr>
<tr>
<td>9)</td>
<td>VA 1 LX(2,1) LX(4,2) LX(6,3)</td>
</tr>
<tr>
<td>10)</td>
<td>FI KA(1) KA(2)</td>
</tr>
<tr>
<td>11)</td>
<td>CO LX(6,1)=TX(4)</td>
</tr>
<tr>
<td>12)</td>
<td>CO LX(6,2)=TX(2)</td>
</tr>
<tr>
<td>13)</td>
<td>CO TD(6,2)=TX(4)*TD(2,2)</td>
</tr>
<tr>
<td>14)</td>
<td>CO TD(6,4)=TX(2)*TD(4,4)</td>
</tr>
<tr>
<td>15)</td>
<td>CO PH(3,3)=PH(1,1)*PH(2,2)+PH(2,1)**2</td>
</tr>
<tr>
<td>16)</td>
<td>CO TD(6,6)=TX(2)**2<em>TD(4,4)+TX(4)**2</em>TD(2,2)+PH(1,1)*TD(4,4) C +PH(2,2)*TD(2,2)+TD(2,2)*TD(4,4)</td>
</tr>
<tr>
<td>17)</td>
<td>CO KA(3)=PH(2,1)</td>
</tr>
<tr>
<td>18)</td>
<td>CO TX(6)=TX(2)*TX(4)</td>
</tr>
<tr>
<td>19)</td>
<td>PD</td>
</tr>
<tr>
<td>20)</td>
<td>OU AD=OFF IT=100</td>
</tr>
</tbody>
</table>

a. The KA designation represents the vector of latent variable means.

b. This fixes to zero the covariances between the latent product and the latent variables that make up the latent product. These values should be near zero as a result of centering. This step is not crucial and can be omitted.

c. LX(2,1) and LX(4,2) are fixed at 1 to define the scales. LX(6,3) carries the path from the latent product to its indicator and is fixed at one because this coefficient must be estimated indirectly through the \( \tau_{y} \) and \( \tau_{x} \) values. It should also be noted that relationships with \( y \) although typically contained in the gamma matrix linking exogenous to endogenous concepts, are here contained in the lambda-x matrix because of the "all-x" nature of this model. This also means that the first value in the tau-x vector pertains to \( y \), not to \( x1 \). Thus, \( \tau_{y} \) and \( \tau_{x} \) pertain to the second and fourth values in the tau-x vector, which are \( x1 \) and \( x3 \), the variables that make up the indicator of the latent product. We mention this to preempt confusion from the wording in Jöreskog and Yang.

d. These values are set to zero as a result of centering.

e. Lines 11 and 12 serve to set the loadings of the indicator of the latent product onto the main effect latent variables equal to the relevant intercept, as mentioned above. Again, the purpose of this is to allow estimation of the loading of the indicator of the latent product onto the latent product. Lines 13 and 14 allow the error associated with the indicator of the latent product to correlate with the errors associated with the variables that comprise the product.

g. This statement serves to set the variance of the latent product equal to the product of the variances of its components plus the square of their covariance, as per Hayduk (1987, Eq. 7.44).

h. The expected value of the latent product is equal to the covariance between its components.

i. The expected value of the observed product is equal to the product of the expected values of the components of the product.
17 in Table 4, cannot be represented in SIMPLIS language at this time. Thus, only matrix language versions of the Jöreskog and Yang procedure are possible.

**Ping (1995)**

Another approach involving a single indicator of the latent product is described by Ping (1995). Ping suggested that the product of the sums of the relevant indicators be used as the sole indicator of the latent product. For example, suppose that two latent variables \(X\) and \(Z\), with indicators \(x_1, x_2\) and \(z_1, z_2\), respectively, are hypothesized to interact in their effect on a third latent variable, \(Y\), which is indicated by a single observed variable \(y\). Ping suggested that the computed variable \([ (x_1 + x_2) \times (z_1 + z_2) ]\) be used as the indicator of the latent product.

The loading and error for the indicator of the latent product are given by the following equations (Equations 4 and 5 from Ping, 1995):

\[
\lambda_{xz} = (\lambda_{x1} + \lambda_{x2})(\lambda_{z1} + \lambda_{z2})
\]

\[
\theta_{xz} = (\lambda_{x1} + \lambda_{x2})^2\text{VAR}(X)(\theta_{e1} + \theta_{e2}) + (\lambda_{z1} + \lambda_{z2})^2\text{VAR}(Z)(\theta_{e1} + \theta_{e2}) + (\theta_{e1} + \theta_{e2})(\theta_{e1} + \theta_{e2})
\]

Because the values on the right side of these equations are available from the additive version of the measurement model for Figure 7, Ping (1995) recommended that the additive model be established first. The relevant values from this analysis can then be used to fix the paths associated with the latent product in the multiplicative model.
Anderson and Gerbing (1988) pointed out that the fixing of certain parameter values in a structural model based on estimates from the measurement model is perfectly justified when the latent variables are unidimensional. In the above example, if $\xi_1$ and $\xi_2$ are unidimensional, then the paths and values associated with their indicators are unaffected by the presence of other variables in the model. Thus, the $\lambda$, $\theta$, and $\Phi$ values from the additive model can be plugged into Equations 12 and 13, and the resulting values can be used to fix the corresponding values associated with the indicator of the latent product in the structural model. Of course, to the extent that unidimensionality cannot be assumed, this two-step procedure is likely to be problematic. Nevertheless, consider the example used previously to illustrate the Jöreskog and Yang (1996) procedure. This example contained two indicators for each of two latent variables and a single indicator of the dependent variable. After centering, the first step in the Ping (1995) procedure would be to compute variables that represent the sums of the indicators of each latent variable that is to go into the latent product. Once this has been done, the product of these summed variables can be computed. This product variable will serve as the indicator of the latent product. The next step would be to estimate the values associated with the additive measurement model. As per Equations 12 and 13, the $\lambda$, $\theta$, and variance values from this analysis can be used to compute the $\lambda$ and $\theta$ values for the indicator of the latent product. These values can then be fixed in the test of the multiplicative model.
Tables 5 and 6 contain matrix and SIMPLIS language versions of the Ping (1995) two-step procedure. It should be noted, however, that it is also possible to produce a single-step, matrix language version of the Ping procedure. Specifically, if data are centered prior to analysis and if latent main effect variables are reasonably unidimensional, then the path coefficients associated with the additive portion of the model should be relatively unaffected by the presence of product terms. Thus, the values from the additive model that are to comprise the path coefficient linking the latent product to its indicator need not be generated in a separate step. Instead, matrix command lines such as those presented in Table 7 can be used in a single step. Unfortunately, the single-step procedure could not be conducted using SIMPLIS command lines for the same reasons that the Jöreskog and Yang (1996) procedure could not be represented with SIMPLIS command lines.

Of particular note in the command lines in Tables 5 through 7 are (a) the lack of use of indicators to define the scales of the latent variables, although there is no reason to exclude such operations; (b) the use of Ping (1995) Equations 4 and 5 to set loading and error variance values for the indicator of the latent product (lines 10-11 and 12-13 of Step 2 in Table 5; lines 8 and 10 in Step 2 of Table 6; and lines 9 and 10 in Table 7). Other command lines serve purposes similar to those served by corresponding lines in the Jaccard and Wan (1995) and Jöreskog and Yang (1996) procedures.
Table 2 contains results for both the one-step and two-step versions of Ping (1995). The small differences between the two sets of output are likely due to the fact that in the two-step version, values for Equations 12 and 13 are taken from a main-effects-only model, whereas, in the one-step version, values are taken from a main-effects-plus-products model. As was mentioned earlier, the presence of the products should make little difference, and indeed, such is the case here.

As can be seen from Table 2, the one and two-step versions of the Ping (1995) procedure produce results that are similar to one another and to the results from the Jaccard and Wan (1995) and Jöreskog and Yang (1996) procedures. It should be noted that the standardized values for the loading of the sexual harassment experiences variable onto the latent product in the two runs based on Ping were almost identical (.23 and .26, respectively) as were the values for the loading of the sexual harassment experiences variable onto the supervisory support latent variable (−.70 and −.71, respectively). The only value from sections 3a and 3b in Table 2 that requires some mention is the variance value for the latent product (PH(3,3)) from the one-step version. As was mentioned earlier, this value should be different from 1 and can be estimated as the product of the variances of the latent main effect variables plus the square of their

### Table 6
LISREL SIMPLIS Code for the Ping (1995) Procedure (two steps)

<table>
<thead>
<tr>
<th>Step 1: Initial run: Model excludes latent product</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Observed Variables: Y X1 X2 Z1 Z2</td>
</tr>
<tr>
<td>2) Covariance Matrix From File Ping95.COV</td>
</tr>
<tr>
<td>3) Sample Size 300</td>
</tr>
<tr>
<td>4) Latent Variables: Climate Super</td>
</tr>
<tr>
<td>5) Relationships:</td>
</tr>
<tr>
<td>6) X1 X2 = Climate a</td>
</tr>
<tr>
<td>7) Z1 Z2 = Super a</td>
</tr>
<tr>
<td>8) Y = Climate Super</td>
</tr>
<tr>
<td>9) LISREL OUTPUT: AD=OFF IT=100</td>
</tr>
<tr>
<td>10) End of Problem</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2: Model includes latent product and set values from Step 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Observed Variables: Y X1 X2 Z1 Z2 XZ</td>
</tr>
<tr>
<td>2) Covariance Matrix From File Ping95.COV</td>
</tr>
<tr>
<td>3) Sample Size 300</td>
</tr>
<tr>
<td>4) Latent Variables: Climate Super ClimXSup</td>
</tr>
<tr>
<td>5) Relationships:</td>
</tr>
<tr>
<td>6) X1 X2 = Climate a</td>
</tr>
<tr>
<td>7) Z1 Z2 = Super a</td>
</tr>
<tr>
<td>8) XZ = 2.396*ClimXSup</td>
</tr>
<tr>
<td>9) Y = Climate Super ClimXSup</td>
</tr>
<tr>
<td>10) Set the Error Variance of XZ to 6.350</td>
</tr>
<tr>
<td>11) Set the Correlation Climate-ClimXSup to 0</td>
</tr>
<tr>
<td>12) Set the Correlation Super-ClimXSup to 0</td>
</tr>
<tr>
<td>13) LISREL OUTPUT: AD=OFF IT=100</td>
</tr>
<tr>
<td>14) PD</td>
</tr>
<tr>
<td>15) End of Problem</td>
</tr>
</tbody>
</table>

a. Ping’s (1995) procedure does not involve the fixing of the first indicator of $$\xi_1$$ and $$\xi_2$$ to define their respective scales.

b. This value is obtained from Ping (1995, Eq. 4).

c. This value is obtained from Ping (1995, Eq. 5).
covariance \((PH(1,1) \cdot PH(2,2) + PH(2,1)^2); \) Bornstedt & Goldberger, 1969). Although not called for by Ping, \(PH(3,3)\) can simply be fixed at \(PH(1,1) \cdot PH(2,2) + PH(2,1)^2\), which, in the present case, would be \(1 \cdot 1 + .64^2 = 1.41\). Alternatively, the variances of the latent main effect variables could be constrained to equal those of their indicators, as is done in the other procedures described in this article. These values could then be used to compute the variance of the latent product as per the above equation from Bornstedt and Goldberger (1969).

Figure 8 contains a plot of the interaction using the values from the two-step procedure. As can be seen, this plot is almost identical to those from the Jaccard and Wan (1995) and Jöreskog and Yang (1996) procedures.

**Mathieu et al. (1992)**

A fourth method for testing interactions in SEM comes from Mathieu et al. (1992). This procedure is similar to that described in Ping (1995) in that measurement properties established in initial steps are used to fix values in the structural model. Consider once again the example used to illustrate the Jöreskog and Yang (1996) procedure. The first part of the Mathieu et al. procedure involves creation of composites for each of the latent variables \((\xi_1 \text{ and } \xi_2)\) that are to constitute the latent product by summing the indicators of each of these component variables and standardizing (which includes centering) each of these composites. Let us call these composites \(Z\xi_1\) and \(Z\xi_2\). Second, these standardized scale scores are multiplied together to form the “latent” product, \(\xi_3\). Third, the measurement properties for \(Z\xi_1\) and \(Z\xi_2\) are fixed using the square roots of the scale reliabilities. Specifically, the \(\lambda\) values relating the latent variables \(\xi_1\) and \(\xi_2\) to their indicator variables are set equal to the square roots of the reliabilities of \(Z\xi_1\).
and $Zx_2$, and the $\theta$ values for each of these observed variables are set equal to the product of its variance and one minus its reliability (Jöreskog & Sorbom, 1993a). With these values fixed, the additive model is then tested for the purpose of discovering the correlation between the latent variables $\xi_1$ and $\xi_2$ (see Figure 9).
Fourth, the values from the analysis of the additive model are used to compute the reliability for the product term using the following formula from Bornstedt and Marwell (1978):

\[
r_{\xi_1, \xi_2} = \left[ (r_{\xi_1 \cdot \xi_1} * r_{\xi_2 \cdot \xi_2}) + r_{\xi_1 \cdot \xi_2}^2 \right] / (1 + r_{\xi_1 \cdot \xi_2}^2),
\]

where \( r_{\xi_1, \xi_2} \) is the reliability of the product, \( r_{\xi_1 \cdot \xi_1} \) and \( r_{\xi_2 \cdot \xi_2} \) are the reliabilities of the components of the product, and \( r_{\xi_1 \cdot \xi_2}^2 \) is the square of the correlation between the components of the product. This value can then be used to fix the \( \lambda \) value for the path from the latent product to its indicator in the analysis of the structural model. As with the main effect indicators, the \( \theta \) value for the indicator of the latent product is set equal to the product of its variance and one minus its reliability. The final step is to test the model with and without the path from the latent product to the criterion variable, thus allowing a \( \chi^2 \) test of the difference in fit between the two models. The model that would be tested is depicted in Figure 10. The command lines for matrix and SIMPLIS versions of both the additive and multiplicative portions of this analysis with the data used to illustrate Jöreskog and Yang (1996) are presented in Tables 8 and 9.

Of particular note in Tables 8 and 9 are (a) the fixing of loadings as square roots of reliabilities (lines 6-9 in Table 8 and lines 6-8 in Table 9), and (b) the fixing of error variances as observed variance times 1 minus reliability (lines 11-14 in Table 8 and
Table 8

1) DA NI=4 NO=300
2) LA
3) Y ZX1 ZX2 ZX1X2
4) CM Fl=MATHIEU.COV
5) MO NX=4 NK=3 PH=SY LX=SY
6) Fl LX(2,1) LX(3,2) LX(4,3)\(^a\)
7) VA .819 LX(2,1)
8) VA .921 LX(3,2)
9) VA .800 LX (4,3)
10) FR LX(1,1) LX(1,2) LX(1,3)
11) FI TD(2) TD(3) TD(4)\(^b\)
12) VA .329 TD(2)
13) VA .151 TD(3)
14) VA .432 TD(4)
15) FR TD(1)
16) FI PH(3,1) PH(3,2)
17) VA 0 PH(3,1) PH(3,2)\(^c\)
18) PD
19) OU AD=OFF IT=100

\(^a\) Lines 6-9 fix the paths from the latents to the indicators at the square roots of the reliabilities.
\(^b\) Lines 11 through 14 fix error variances equal to observed variance times 1 minus the reliability.
\(^c\) These values should be near zero as a result of centering.

Table 9
LISREL SIMPLIS Code for the Mathieu, Tannenbaum, and Salas (1992) Procedure

1) Observed Variables: Y ZX1 ZX2 ZX1ZX2
2) Covariance Matrix From File MATHIEU.COV
3) Sample Size 300
4) Latent Variables: Climate Super ClimXSup
5) Relationships:
6) ZX1 = .819*Climate\(^a\)
7) ZX2 = .920*Super
8) ZX1ZX2 = .800*ClimXSup
9) Y = Climate Super ClimXSup
10) Set the Error Variance of ZX1 to .329\(^b\)
11) Set the Error Variance of ZX2 to .151
12) Set the Error Variance of ZX1ZX2 to .432
13) Set the Correlation Climate-ClimXSup to 0\(^c\)
14) Set the Correlation Super-ClimXSup to 0\(^c\)
15) PD
16) LISREL OUTPUT: AD=OFF IT=100

\(^a\) Lines 6 through 8 fix the paths from the latents to the indicators at the square roots of the reliabilities.
\(^b\) Lines 10 through 12 fix error variances equal to observed variances times 1 minus the reliabilities.
\(^c\) These values should be near zero as a result of centering.

lines 10-12 in Table 9). Other command lines serve purposes similar to those served by corresponding lines in the procedures illustrated previously.
Table 2 contains the results of the LISREL analysis of the sexual harassment data using the Mathieu et al. (1992) procedure. Most of these values are similar to those generated by the previously described procedures. In particular, the weight for the latent product is very close to the others reported in Table 2. However, one difference between the Mathieu et al. values and corresponding values from the other procedures stands out. In all of the other procedures, the climate variable received a small weight, whereas the supervision variable received a relatively large, negative weight. In the Mathieu et al. procedure, it was the climate variable that received the large negative weight. This was also true of the multiple regression analysis and would suggest a somewhat different conclusion. Specifically, whereas the other procedures discussed here suggest that supervisory support is negatively related to sexual harassment experiences but that this negative relationship is attenuated by a positive climate, the results from the Mathieu et al. procedure suggest that climate is negatively related to sexual harassment experiences but that this negative relationship is attenuated by supervisory support. Note that both conclusions suggest that the slope of the line representing the relationship between one of the predictors and the criterion increases as a function of the other predictor. Figure 11 contains a plot of the interaction from the Mathieu et al. procedure.

We do not know why this difference exists. One should keep in mind, however, that standard errors for any model coefficients increase with collinearity. The strong relationship between the climate and support variables is bound to result in relatively high levels of sampling error for coefficients relating these variables to others. Nevertheless, future research should consider the possibility that this difference in weights had a substantive basis.

The Mathieu et al. (1992) approach is very similar to the Ping (1995) approach. Both require the assumption of unidimensionality of the latent variables (Anderson & Gerbing, 1988), both involve the summing of indicators, and both involve the estimation of measurement model values prior to estimation of structural coefficients. The primary difference between the two procedures is that the Mathieu et al. procedure uses formulas taken directly from classical test theory to estimate the values associated with the product term. The Mathieu et al. approach is, therefore, more straightforward for those who have been trained in traditional psychometrics. The two procedures should recover parameters similarly, and given that Ping found that his approach recovered parameter values almost as well as the full Kenny and Judd (1984) procedure, the Mathieu et al. approach is likely to produce accurate parameter estimates as well. Also, because only one indicator for the latent product is used in the Mathieu et al. approach, the calculations are relatively straightforward, and the procedure can be implemented using either matrix or SIMPLIS language, the Mathieu et al. approach is one of the more user friendly of the available approaches.

Ping (1996b)

One final procedure worth mentioning is that described in Ping (1996b). This procedure is simply a two-step version of the Jaccard and Wan (1995) procedure in which certain values in a second run are fixed based on values from a first, additive run. The advantage that this procedure has over the Jaccard and Wan procedure is that it can be implemented in the SIMPLIS language. This is due to the fact that the complex products that comprise the Θ (TD) values attached to the indicators of the latent product
(lines 17-20 in Table 3) are computed outside of LISREL. The values that result are then inserted in a second step (lines 22-25 of Table 10).

Of particular note in Table 10 are (a) the constraining of the path from the latent product to $X_2 Z_2$ equal to the product of loading of $X_2$ onto climate and the loading of $Z_2$ onto supervision from the first step (line 13 in Step 2), and (b) the fixing of values relating to the latent product and its indicators, as was done in the Jaccard and Wan (1995) procedure.

Table 2 contains the results from the analysis of the sexual harassment data using the Ping (1996b) procedure, and Figure 12 contains a plot of the interaction. There is little that needs to be said about these results, as they are almost identical to those from the Jaccard and Wan (1995) procedure. The differences that exist are simply due to the fact that in the Ping (1996b) procedure, values from an additive-model-only run are used to fix values in the full model, whereas in the Jaccard and Wan procedure, values from the full model are used.

### Summarizing the Results From the Illustrations

In the above section, five procedures for testing multiplicative models in LISREL were illustrated using data relating sexual harassment experiences to climate for harassment and supervisory support. Results from the Jaccard and Wan (1995), Jöreskog and Yang (1996), and Ping (1995, 1996b) procedures were very similar, suggesting a main effect for supervisory support and an interaction between climate and supervisory support such that a supportive supervisor can attenuate the negative influence of a hostile climate. Unstandardized coefficients from these procedures for the main effect of supervisory support ranged from $-.54$ to $-.72$, and unstandardized coef—
TABLE 10
LISREL SIMPLIS Code for the Ping (1996b) Procedure (two steps)

Step 1: Initial run: Model excludes latent product

1) Observed Variables: Y X1 X2 Z1 Z2
2) Covariance Matrix From File J&W95.COV
3) Sample Size 300
4) Latent Variables: Climate Super
5) Relationships:
6) X1 = 1*Climate
7) X2 = Climate
8) Z1 = 1*Super
9) Z2 = Super
10) Y = Climate Super
11) LISREL OUTPUT: AD=OFF IT=100
12) End of Problem

Step 2: Model includes latent product and set values from Step 1

1) Observed Variables: Y X1 X2 Z1 Z2 X1Z1 X1Z2 X2Z1 X2Z2
2) Covariance Matrix From File J&W.COV
3) Sample Size 300
4) Latent Variables: Climate Super ClimXSup
5) Relationships:
6) X1 = 1*Climate
7) X2 = Climate
8) Z1 = 1*Super
9) Z2 = Super
10) X1Z1 = 1*ClimXSup
11) X1Z2 = ClimXSup
12) X2Z1 = ClimXSup
13) X2Z2 = .952* ClimXSup\(^a\)
14) Y = Climate Super ClimXSup
15) Set ClimXSup- >X1Z2 Equal to Super- >Z2
16) Set ClimXSup- >X2Z1 Equal to Climate- >X2
17) Set the variance of ClimXSup to .532\(^b\)
18) Set the Error Covariance Between X1Z1 and X1Z2 Free\(^c\)
19) Set the Error Covariance Between X1Z1 and X2Z1 Free\(^c\)
20) Set the Error Covariance Between X1Z2 and X2Z2 Free\(^c\)
21) Set the Error Covariance Between X2Z1 and X2Z2 Free\(^c\)
22) Set the Error Variance of X1Z1 to .555\(^d\)
23) Set the Error Variance of X1Z2 to .911
24) Set the Error Variance of X2Z1 to .780
25) Set the Error Variance of X2Z2 to 1.424
26) Set the Correlation Climate-ClimXSup to 0
27) Set the Correlation Super-ClimXSup to 0
28) PD
29) LISREL OUTPUT: AD=OFF IT=100

---

a. Line 13 constrains the path from the latent product to X2Z2 equal to the product of climate→X2 and super→Z2. The values for this operation are obtained from the initial run (Step 1).

b. Line 17 fixes the variance of the latent product equal to the product of the variances of its components plus the square of their covariance, as per Hayduk (1987, Eq. 7.44). The values necessary for this operation are obtained from the initial run (Step 1).

c. Errors for the product indicators should correlate with each other because the indicators share components.

d. Lines 22 through 25 constrain the variances of the indicators of the latent product as per Equations 7 through 13 in Jaccard and Wan (1995). The values for these formulas are obtained by conducting an initial run (Step 1).
Coefficients from these procedures for the interaction effect ranged from .14 to .27. The range for standardized values is even smaller (−.65 to −.71 and .21 to .26, respectively). The Mathieu et al. (1992) procedure suggested slightly different conclusions. Although the latent product in the Mathieu et al. procedure also received a significant positive weight (unstandardized = .17, standardized = .24), it was the climate variable instead of the supervision variable that had the significant main effect. Because of the sizable relationship between the climate and supervision variables, a considerable amount of sampling error would exist for the model coefficients relating these latent variables to other variables (see Cohen & Cohen, 1983, chap. 3, for a discussion of the effects of collinearity on the sampling distribution of model coefficients). Nevertheless, future research should investigate the reasons that the Mathieu et al. procedure might produce results that differ from those produced by other procedures.

With regard to fit statistics, the Jaccard and Wan (1995), Ping (1995), Mathieu et al. (1992), and Ping (1996b) procedures produced similar, encouraging fit values, although the root mean square error of approximation for the Mathieu et al. procedure was a bit larger than it was for the other procedures. The Jöreskog and Yang (1996) procedure resulted in even more encouraging fit values, although this may have been due to the lack of overidentifying restrictions relative to the Jaccard and Wan and both Ping procedures.

**Figure 12:** Plot of the Interaction Between Supervisor Support and Sexual Harassment Climate Using the Ping (1996b) Procedure

*Figure 12: Plot of the Interaction Between Supervisor Support and Sexual Harassment Climate Using the Ping (1996b) Procedure*

Summarizing Best Practices

Before concluding, it might be helpful to take a more holistic view of the issues raised here. An MSEM analysis involves many decision points, and it is these decision
points with which we have concerned ourselves. Let us then review the issues raised in this article as they relate to the decisions that one makes in planning and conducting an MSEM analysis.

**Review of Decisions**

We begin our review by assuming that adequate data have already been collected.

**Centering versus not:** Centering can often be helpful. Although some authors express concern over the potential biasing effects of centering in structural equation models, there is little empirical research supporting these concerns. Centering is particularly useful when degrees of freedom are likely to be an issue because centering may allow paths between products and their components to be omitted. Such an omission is certain to reduce model fit, but the reduction will be small if violation of multivariate normality is not great, and the gain in degrees of freedom can sometimes make the difference between over- and underidentification.

**Modeling the product term.** Because a multiplicative term is a tool through which variables of interest to the social/organizational sciences are evaluated as opposed to a variable of interest in and of itself, there is little need to consider any links involving the multiplicative term other than those relating to the particular interaction that it was created to test.

**“Measuring” the product term.** Although some have suggested that all possible cross products of observed variables be used as indicators of the latent product, this approach can create a variety of problems. Given these problems and the relative lack of advantages associated with a large number of indicators for a latent product, a single indicator is usually sufficient.

**Choosing an estimator.** ML is not a panacea, but it appears to work reasonably well in many situations. Its assumption of multinormality is typically violated in MSEM, but unless the violation is egregious and the structural errors behave strangely (i.e., they correlate with exogenous indicators or are themselves nonnormally distributed), the desirable properties of the ML estimator hold. Estimators based on asymptotic distribution-free theory are ill advised with sample sizes smaller than 1,000. The “scaled” estimator suggested by Satorra and Bentler (1988) appears to be a reasonable alternative.

**Choosing a method.** The available methods vary with respect to technical elegance and usability. The Jöreskog and Yang (1996) and Jaccard and Wan (1995) procedures are among the more elegant. They are also among the more complicated, both conceptually and operationally. Procedures suggested by Ping (1995, 1996) seem to be more user friendly and recover parameter values well. The procedure suggested by Mathieu et al. (1992) is the simplest to implement, and it is likely to be the easiest to understand for those trained in classical test theory. One option is to begin with one of the more elegant procedures. If one finds it to be manageable and if it converges, then one need not look any further. If the more elegant procedures prove too unwieldy or if they fail to converge, then one might turn to a more user-friendly procedure. For example, the more user-friendly approaches, particularly the Mathieu et al. procedure, may be espe-
cially useful when testing more complicated theoretical models that include both mediated and moderated relationships (e.g., see Mathieu et al., 1992).

### Discussion

The purpose of this article was to review the techniques available for testing multiplicative effects in structural equation models. The techniques developed by Kenny and Judd (1984), Jöreskog and Yang (1996), Ping (1995), Mathieu et al. (1992), Jaccard and Wan (1995), and Ping (1996) were described, and the LISREL code necessary to conduct these procedures was offered. Although all of these procedures are likely to recover parameter values (indeed, results were similar across the different procedures for our data), the Mathieu et al. and Ping (1995) procedures are more straightforward conceptually and operationally. This is likely to make these procedures the easiest to implement and the least likely to produce problems with convergence.

As understanding of the original work of Kenny and Judd (1984) has spread and as existing programs have been developed to accommodate the requirements of models with multiplicative effects, more procedures for testing such models have been offered. Table 11 distinguishes between these procedures with respect to transportability to SIMPLIS, need for external calculations, need for multiple LISREL steps, and number of latent product indicators.

More work needs to be done to delineate the advantages and disadvantages of these procedures. In particular, simulations that examine when and why (and if) the different procedures produce different results should be conducted. For example, it may be that the different approaches for forming the latent product respond differently to characteristics of main effect indicators such as their distributional properties or their relationships with one another. Regardless of the questions that are explored with such studies, we suggest that this work should not focus solely on comparisons with respect to parameter recovery. As was mentioned above, a tool that is elegant yet too cumbersome for most potential consumers to use is a tool of questionable value. Even if the tool were somewhat less precisely functional, it might, nevertheless, be of greater value if it were less cumbersome. It is important to consider the fact that methods differ

### Table 11

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Computes/ Constraints</th>
<th>External Calculations</th>
<th>Single Indicator for Latent Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathieu, Tannenbaum, and Salas (1992)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Ping (1995)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Ping (1996)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Jaccard and Wan (1995)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jöreskog and Yang (1996)</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Only if the two-step procedure is used.
not only in terms of quantitative elegance but also in terms of usability. When one of these dimensions is ignored in favor of the other, we often end up with tools for which there is no use.

**Notes**

1. Thanks to an anonymous reviewer for raising this issue.
2. Thanks to an anonymous reviewer for this suggestion.
3. Although no empirical work has been done to establish the accuracy with which the Mathieu, Tannenbaum, and Salas (1992) procedure recovers parameter values, the similarity of the procedure with that of Ping (1995) allows us to extrapolate the results of Ping to the Mathieu et al. procedure.
4. Thanks to an anonymous reviewer for pointing this out.

**References**


**Jose M. Cortina** is an associate professor in George Mason University’s Industrial/Organizational Psychology program. He received his Ph.D. in 1994 from Michigan State University. His research interests include everything unrelated to groups and teams.

**Gilad Chen** is an assistant professor in the Industrial/Organizational Psychology program at Georgia Institute of Technology. He received his Ph.D. in 2001 from George Mason University, despite Jose Cortina’s objections. His research interests include work motivation, leadership, and team processes, and research methods and techniques.

**William P. Dunlap** is a professor in the Department of Psychology at Tulane University. He received his Ph.D. a long, long time ago from Tulane. He has published papers on a wide variety of methodological and statistical topics.
Single-Item Reliability: 
A Replication and Extension

JOHN P. WANOUS  
The Ohio State University 

MICHAEL J. HUDY  
SHL USA, Inc.

The reliability of a single-item measure of student-rated college teaching effectiveness was estimated with two different methods and at two levels of analysis. The two methods are the correction for attenuation formula and factor analysis. The two levels of analysis are the group level (10,682 classes) and the individual level (323,262 students). Reliability estimates were higher using factor analysis (.88) than the correction for attenuation formula (.64), and they were higher using group-level data (.82) than individual-level data (.70). Based on the assumptions and limitations of each method used, the authors conclude that a minimum estimate of .80 for single-item reliability is reasonable for group-level data. The authors reaffirm a minimum reliability estimate of .70 for individual-level data, as previously concluded by Wanous, Reichers, and Hudy, who estimated single-item reliability for measures of overall job satisfaction using individual-level data.

Despite their frequent use as measures of overall job performance, overall job satisfaction, intention to quit, whether to hire, whether to promote, and so on, single-item measures are often discouraged for the conduct of scholarly research. At first, this appears to be an unfortunate conflict between common human resource practice in many organizations versus the recommendations of some researchers and journal editors. This particular conflict may also be representative of the general tension between the science and the practice of psychology. (See Rice, 1997, for a discussion of this tension).

Editors of scholarly journals and those who review for them tend to be very concerned with measurement reliability and frequently treat it as a necessary condition for the acceptance of research for publication. In the case of single-item measures, the emphasis on measurement reliability has caused problems for those using them in research. First, it is widely believed that the reliability of a single-item measure cannot be estimated. Second, it is commonly believed that single-item reliability would be unacceptably low, if it could be estimated.

Recently, however, work on single-item measures of overall job satisfaction has challenged both of these assumptions about single-item reliability. Wanous and
Reichers (1996) first showed how the minimum level of reliability for a single-item measure could be estimated by using the well-known correction for attenuation formula. Following this, it was concluded that a minimum reliability estimate of .70 for single-item measures of overall job satisfaction was reasonable (Wanous, Reichers, & Hudy, 1997). The conclusion about single-item reliability was based on a meta-analysis (17 samples, 28 correlations, 7,682 persons) of the accumulated research using single-item measures of job satisfaction.

The purpose of the present study is to replicate and extend the research about single-item reliability. There are four ways that the present research adds to the small body of existing research. First, a measure somewhat different from job satisfaction is examined here, that is, a single-item measure of overall teaching effectiveness of college faculty as rated by their students. This is a different type of concept/measure than job satisfaction but is complementary because job satisfaction and work role effectiveness have been among the most studied factors in research on people at work. The present study of teaching effectiveness is different from research on job satisfaction because it concerns a rating about someone else’s behavior rather than a self-rating of one’s own internal state. Thus, examining teaching effectiveness expands the criterion domain in the continuing investigation of single-item reliability.

Second, the particulars of the data reported here are such that we believe a better estimate of single-item reliability may be obtained here, as compared with the previous research by Wanous and colleagues. There are three reasons for this belief: (a) The same measure is used throughout, which eliminates one source of between-studies variance; (b) the present sample is probably more homogeneous than found in the typical meta-analysis of many different samples, and this reduces another potential source of between-studies variance; and (c) the size of the present sample is much larger than typically found in organizational research, including most meta-analyses as well.

Third, the data are analyzed at two levels of analysis: individual and group. Although the appropriate level of analysis for the data reported here is the group level, we also analyze the same data at the individual level for comparison purposes. This is because the previously reported research on single-item reliability involved data from the individual level of analysis (Wanous et al., 1997; Wanous & Reichers, 1996). Thus, this is the first study of single-item reliability to use group-level data.

Fourth, a second method of estimating single-item reliability is used in addition to the correction for attenuation formula previously introduced by Wanous and colleagues (Wanous et al., 1997; Wanous & Reichers, 1996). This second method is based on factor analysis. According to the theory of factor analysis, the communality of the single-item measure of overall teaching effectiveness in a factor analysis with other teaching-effectiveness-related items can be interpreted as a second way to estimate reliability (Harman, 1967). Each of these methods for single-item reliability estimation is explained next.

**Estimating Reliability: The Correction for Attenuation Formula**

The well-known correction for attenuation formula can be found in all textbooks on psychometrics (e.g., Nunnally & Bernstein, 1994, p. 257). It is expressed as the following:
\[ r_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \]

where \( r_{xy} \) = the observed correlation between variables \( x \) and \( y \), \( r_{xx} \) = the reliability of variable \( x \), \( r_{yy} \) = the reliability of variable \( y \), and \( r_{xy} \) = the assumed underlying construct correlation between \( x \) and \( y \).

This formula is most commonly applied to situations in which variables \( x \) and \( y \) come from two different domains, for example, job satisfaction and job performance, and so on. However, the formula can also be applied to a situation in which both variables come from the same conceptual domain. When this occurs, Nunnally (1978, p. 220) said that

the correlation between two such tests would be expected to equal the product of the terms in the denominator and consequently \( r_{xy} \) would equal 1.00. . . If \( r_{xy} \) were 1.00, \( r_{xy} \) would be limited only by the reliabilities of the two tests:

\[ r_{xy} = \sqrt{r_{xx} r_{yy}}. \]

Estimating Reliability: Factor Analysis

Factor analysis can be used as a second method for estimating single-item reliability (Weiss, 1976). According to Harman (1967, pp. 16-19), the variance of an item can be represented by the following formula: Total Variance = Communality + Specificity + Unreliability. Thus, the reliable variance for an item is the sum of its communality and its specificity, which means that the communality is less than or equal to the reliability of an item. When there is no specific variance, the communality equals the reliability. Thus, communality can be considered a conservative estimate of single-item reliability. This method has been used at least once before (although it was not the main point of that research and is quite easy to overlook) in a monograph by Arvey, Landon, Nutting, and Maxwell (1992, p. 1000).

Method

Sample

Student ratings were collected over a 2-year period at a very large public university. These ratings were collected at the main campus and five branch campuses located throughout the state and included both undergraduate and graduate student classes. Classes were taught by instructors ranging in rank from graduate teaching associate to full professor. Classes with fewer than five student ratings of teaching effectiveness were excluded because of their small size. The primary concern with these very small classes is that a single student’s rating can be quite significant on the class mean rating of the instructor.

All analyses were conducted at two levels of analysis: the group level (\( N = 10,682 \) classes) and the individual level (\( N = 323,262 \) student ratings). The group-level data are the relevant data for this particular study because only the class means for the student ratings of teaching effectiveness are used by the university, as is typical for stu-
dent ratings. The individual-level data are similar (in terms of level of analysis) to those reported by Wanous et al. (1997).

Using class mean ratings is consistent with the intended purpose of the Student Evaluation of Instruction (SEI) instrument, that is, to provide an overall evaluation of classroom teaching effectiveness for a particular course. When individual student ratings within a class are used as the unit of analysis, variations in ratings reflect individual differences in the perceptions of particular students. By using class mean ratings, these individual differences among students’ perceptions are removed to some extent. According to Cranton and Smith (1990), variations in ratings at the class mean level reflect perceived differences in teaching effectiveness among instructors. The construct of concern here is an instructor’s overall rating for a particular class, which can only be obtained by averaging the ratings made by individual students.

Measures

The SEI was developed by this university in 1994 for the singular purpose of evaluating teaching effectiveness for personnel decisions such as promotion, tenure, and merit pay. It replaced a much longer rating form (developed in 1977) that was designed to provide both evaluative and developmental feedback.

At the same time that students complete the SEI, they are also given blank sheets labeled “Student Comments.” After the completion of the course, these written comments are returned to instructors and represent the primary source of specific, developmental feedback, to the extent that it is provided in these comments. In contrast, the SEI forms are optically scanned and scored. Then, the numerical results are sent both to individual instructors as well as their respective department chairs and college deans.

The SEI was constructed with the joint efforts of five different committees. In all, 43 faculty, staff, and students worked on its development. The development process included face-to-face interviews with faculty and administrators, telephone interviews with faculty and students, a review of the teaching evaluation literature, and two pretests.

The SEI is a one-page form that has nine facets of teaching effectiveness and one overall rating. The nine effectiveness facet items are worded in general terms because they were designed to apply to the entire portfolio of classes at this university. The nine facets are measured with a 5-point agree/disagree scale. In contrast, the overall rating is measured with a 5-point Likert-type scale having two anchors: 5 = excellent and 1 = poor. In addition to these 10 items, other questions on the SEI include class standing (freshman, sophomore, etc.), self-reported cumulative grade point average, and the reason for enrolling in the class (required vs. elective).

Meta-Analysis Procedure

A meta-analysis was conducted (on the group-level data) to increase our confidence in the average correlation between the single-item measure of overall teaching effectiveness and the nine-item scale of teaching effectiveness facets. Prior to our study, research by the university administration found that the mean SEI ratings were affected by two factors: (a) size of class (the smaller the class, the higher the SEI rating) and (b) whether the class was required or an elective (electives had higher SEI ratings).
What we did not know, however, was whether the correlation between the single-item overall rating and the sum of the nine facets would be affected by these, or other, situational factors. Because this particular correlation is the statistic of interest here, we first conducted a meta-analysis to assess the possible effects of situational factors. Given that the mean SEI ratings are affected by two factors, the meta-analysis on the correlation seemed cautious yet desirable.

All meta-analyses were done with the META 5.3 software developed by Schwarzer (1989), which uses the Hunter and Schmidt (1990) method. META 5.3 can accommodate up to 500 effect sizes. Despite its relatively high limit, several random samples of 500 classes had to be drawn because the data pool is extraordinarily large. Because the Schwarzer (1989) meta-analysis software was less familiar to us, we compared its results with those from the Switzer (1991) software, which has been used in several publications (including the predecessor of this one by Wanous et al., 1997) but which has a much lower sample size limit. No differences were found between these two programs.

Separate meta-analyses were conducted for four groups of classes with different class size minimums. The first meta-analysis was run on a random sample of 500 classes selected from the entire sample of classes (10,682). Because this group included relatively small classes (5-19), three additional meta-analyses were conducted with sample-size minimums of 20, 40, and 60, respectively. When the class size minimum is raised to 60, no random sampling was necessary because the number of classes was less than 500. This is a second reason to conduct these meta-analyses, because it is a way to assess the representativeness of the random samples used when the class size limits are lower and 500 had to be drawn from the entire pool of classes (10,682).

Finally, the average correlation between the single-item overall teaching effectiveness measure and the scale of nine teaching effectiveness facet items is corrected for unreliability in the nine-item scale. This is common practice in “bare bones” meta-analysis (Hunter & Schmidt, 1990, p. 156). However, the corrected correlation is used only in the meta-analyses. It is not used in the estimation of single-item reliability because that would result in a double correction and an overestimate (see Wanous et al., 1997).

Factor Analysis Procedure

To begin with, two factor analysis methods were compared, principle axis and maximum likelihood, on the class mean ratings (N = 10,682) using SPSS for Windows (Release 6.1). This was done as a procedural check to see if these two methods produced different results. In fact, the results were almost identical, so we used the principle axis method for estimating reliability at the two levels of analysis.

Results

Descriptive statistics for all 10 SEI items are presented in Table 1. The means, standard deviations, and item intercorrelations were calculated at the class mean level because this is the relevant level for this situation. All correlations are significant (p < .01).
The results of the four meta-analyses are presented in Table 2. For the first group in Table 2 (minimum class size of five ratings), the mean observed correlation between the scale and single item is .79 ($SD = .16$). When this is corrected for unreliability in the nine-item facet scale, the correlation is .84 ($SD = .14$). The amount of between-class variance due to all artifacts is 35%, of which 30% is due to sampling error and 5% is due to measurement error.

Changing the class size minimums to 20, 40, or 60 has virtually no effect on the average correlations or their respective standard deviations. However, the percentage of between-class variance explained by artifacts decreases as the class size minimums increase. This is the natural and expected consequence of restricting the variance in class size when the minimum size is increased. With a more homogeneous set of classes as the minimum size increases, there is simply less between-class sampling error.

The second phase of the meta-analysis was a search for potential moderators. However, no moderators could be detected from the available data: instructor gender, instructor ethnicity (White vs. non-White), class location (main campus vs. branch), class size, course level (sophomore, etc.), rank of the instructor, or whether the course was a requirement or an elective. Finding no moderators further increased our confidence in the average correlation as representative of the entire sample.

The reliability of the single-item measure of teaching effectiveness was first estimated using the correction for attenuation formula. Because the results of the meta-analyses using different class size minimums were so similar, the single-item reliability estimate is based on the results for the first group (minimum class size of five) as perhaps the most representative sample of the entire portfolio of classes taught at this university because it included classes of all sizes (except for those with fewer than five students).
When using the correction for attenuation formula to estimate single-item reliability, the underlying construct correlation between the scale and single-item measure must be assumed and entered into the formula (Nunnally, 1978; Nunnally & Bernstein, 1994; Wanous & Reichers, 1996; Wanous et al., 1997). The most conservative approach is to assume a 1.0 correlation because relaxing this assumption produces higher estimates.

After assuming a construct correlation of 1.0, the other known elements in the formula are determined and the equation solved for its missing value, that is, the single-item reliability. The group-level data are first reported. The reliability of the nine-item facet scale was estimated as .88. The observed average correlation between the scale and single-item measure is .79. The observed correlation is used rather than the corrected correlation because using the corrected correlation would result in an inflated estimate due to double correction (Wanous et al., 1997). Using these numbers in the correction for attenuation formula results in an estimated minimum-level single-item reliability of .71 for the group-level data. If the assumption about the underlying construct correlation is relaxed to .95, then the minimum reliability estimate increases to .79. If it is relaxed further to .90, the minimum reliability estimate increases to .88.

The individual-level data were next used to estimate single-item reliability with the correction for attenuation formula method. For these data, the correlation between the single-item and the sum of the nine facets is .73, and the reliability of the nine-item scale of facets was estimated as .92. This results in .58 as the estimate of the minimum single-item reliability. If the assumption about the underlying construct correlation is relaxed from 1.00 to .95, this estimate increases to .64. If the assumption is relaxed further to .90, the estimate increases to .72.

Next, the reliability of the single-item measure was estimated using principle axis factor analysis. Table 3 shows the results using all 10 SEI items (nine facets and one overall item). The group-level data are shown first.

<table>
<thead>
<tr>
<th>Minimum Size of Class (n)</th>
<th>5</th>
<th>20</th>
<th>40</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample of classes</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>227</td>
</tr>
<tr>
<td>Total number of classes</td>
<td>10,682</td>
<td>3,668</td>
<td>681</td>
<td>227</td>
</tr>
<tr>
<td>Total number of students</td>
<td>9,546</td>
<td>16,220</td>
<td>30,473</td>
<td>20,021</td>
</tr>
<tr>
<td>Weighted mean correlation</td>
<td>.79</td>
<td>.78</td>
<td>.79</td>
<td>.80</td>
</tr>
<tr>
<td>Standard deviation (SD)</td>
<td>.16</td>
<td>.16</td>
<td>.12</td>
<td>.12</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>.53-1.05</td>
<td>.49-1.06</td>
<td>.57-1.02</td>
<td>.59-1.02</td>
</tr>
<tr>
<td>Corrected correlation</td>
<td>.84</td>
<td>.82</td>
<td>.83</td>
<td>.84</td>
</tr>
<tr>
<td>Corrected SD</td>
<td>.14</td>
<td>.15</td>
<td>.12</td>
<td>.11</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>.57-1.11</td>
<td>.53-1.12</td>
<td>.60-1.06</td>
<td>.62-1.07</td>
</tr>
<tr>
<td>Percentage of variance due to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling error</td>
<td>29.86</td>
<td>18.36</td>
<td>14.89</td>
<td>10.61</td>
</tr>
<tr>
<td>Unreliability</td>
<td>5.29</td>
<td>4.21</td>
<td>2.18</td>
<td>2.44</td>
</tr>
<tr>
<td>Total</td>
<td>35.14</td>
<td>22.57</td>
<td>17.07</td>
<td>13.05</td>
</tr>
</tbody>
</table>
For the group-level data, one factor was extracted with an eigenvalue of 7.83, accounting for 78.3% of the variance. Only the communalities are shown because they are directly derived from the factor loadings and because they are the estimates of single-item reliability. The single-item overall effectiveness measure has a communality of .94. This is the highest communality of any item on the SEI, and it can be compared to the average communality of .77 for the nine facet items.

Next, the individual-level data are reported in Table 3. Again, one factor was extracted with an eigenvalue of 6.44, accounting for 64.4% of the variance. The communality of the overall item is .81 and again is the highest communality of any of the items. It can be compared to the average communality of .63 for the nine facets.

Table 4 summarizes the four estimates of single-item reliability based on the two methods and two levels of analysis. Remember that the estimates based on the correction for attenuation formula shown in Table 4 are based on the most conservative assumption about the underlying construct correlation. The reasonableness of this assumption will be discussed later.

**Discussion**

Overall single-item measures have a long history in industrial psychology and human resource management, for example, for hiring, promotion, and termination. They also have a long history in organizational behavior research, for example, appraising one’s overall job satisfaction, assessing one’s intention to quit an organization, or assessing one’s own job performance. Despite their frequent use, it must be remembered that single-item measures are most appropriately used in situations in which the construct of interest is (a) unidimensional rather than multidimensional, (b) clear to the respondents, and (c) sufficiently narrow (Sackett & Larson, 1990).

---

**Table 3**

Principal Axis Factor Analyses: Communalities

<table>
<thead>
<tr>
<th>Item</th>
<th>Class Means (N = 10,682)</th>
<th>Individual Data (N = 323,262)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Organized</td>
<td>.72</td>
<td>.56</td>
</tr>
<tr>
<td>2. Stimulating</td>
<td>.55</td>
<td>.47</td>
</tr>
<tr>
<td>3. Interested</td>
<td>.83</td>
<td>.64</td>
</tr>
<tr>
<td>4. Think</td>
<td>.65</td>
<td>.51</td>
</tr>
<tr>
<td>5. Prepared</td>
<td>.76</td>
<td>.62</td>
</tr>
<tr>
<td>6. Helping</td>
<td>.74</td>
<td>.61</td>
</tr>
<tr>
<td>7. Learned</td>
<td>.90</td>
<td>.75</td>
</tr>
<tr>
<td>8. Atmosphere</td>
<td>.92</td>
<td>.76</td>
</tr>
<tr>
<td>9. Communicated</td>
<td>.85</td>
<td>.72</td>
</tr>
<tr>
<td>Scale average</td>
<td>.77</td>
<td>.63</td>
</tr>
<tr>
<td>Overall rating</td>
<td>.94</td>
<td>.81</td>
</tr>
</tbody>
</table>

*Note.* Organized = the subject matter of the course was well organized; stimulating = the course was intellectually stimulating; interested = the instructor was genuinely interested in teaching; think = the instructor encouraged students to think for themselves; prepared = the instructor was well prepared; helping = the instructor was genuinely interested in helping students; learned = I learned a great deal from this instructor; atmosphere = the instructor created an atmosphere conducive to learning; communicated = the instructor communicated the subject matter clearly; overall rating = overall, I would rate this instructor as . . . .
Summary of Results

The reliability of the single-item measure was estimated in two ways. Furthermore, the data were aggregated at the group level (classes taught) and at the individual level for the sake of comparison. It must be remembered, however, that the group level of analysis is the appropriate one for these data.

The correction for attenuation formula was used as the first method for single-item reliability estimation. For the group-level data, reliability estimates of .71, .79, and .88 were derived, depending on the assumptions made about the underlying construct correlation. For the individual-level data, parallel reliability estimates of .58, .64, and .72 were derived. The lowest of these estimates is based on the most conservative assumption that the underlying construct correlation between the two measures of overall teaching effectiveness is 1.0. As discussed later, it may be reasonable to relax the assumption of a perfect construct correlation to a lower level, such as .95 or .90.

The second method for estimating single-item reliability used factor analysis, in which it is assumed that the total variance of an item has three elements: communality, specificity, and unreliability (Harman, 1967). This means that the reliable variance of an item is equal to the sum of its communality and its specificity. Because of this, communality estimates can be used as a conservative estimate of an item’s reliability. Principle axis factor analysis is reported here, although the maximum likelihood procedure was used on the group-level data for the sake of comparison between methods. Because there were only trivial differences between these two methods of factor analysis, these additional results are not reported but are available.

Differences Between Estimating Methods

The two methods for estimating single-item reliability appear to have somewhat different results if one compares the lowest reliability estimates from the correction for attenuation formula (.70 for the group level and .58 for the individual level) with the estimates from factor analysis (.94 for the group level and .81 for the individual level). However, the two methods probably agree more than is initially apparent because each method of estimating reliability has its own assumptions and limitations. These assumptions and limitations most likely result in those derived from the attenuation formula being underestimates, whereas those estimates derived from factor analysis are most likely overestimates.

First, the reliability estimate based on the correction for attenuation formula may be an underestimate because the assumption of a 1.0 construct correlation is itself too
conservative. For the group-level data, if one assumes that the construct correlation is .95, the estimated minimum reliability increases from .70 to .79. Relaxing this assumption even further to .90 results in a reliability estimate of .88. A similar pattern is found for the individual-level data. The most conservative assumption results in an estimate of .58, but relaxing this results in other estimates of .64 and .72.

It may be reasonable to relax the assumption of a perfect construct correlation somewhat, as was previously done when this method was used to estimate the single-item reliability of overall job satisfaction measures (Wanous et al., 1997). In the case of job satisfaction, the arguments made by Scarpello and Campbell (1983) strongly suggested that the construct correlation should be less than 1.00. Briefly, they argued that any sample of job facet satisfaction items will be deficient when used to measure a global type of construct such as overall job satisfaction. Their data also supported this view. Whether the same can be said for the present case of student ratings of classroom teaching has not yet been determined empirically, although the logical arguments articulated by Scarpello and Campbell very well may apply to this domain too.

A second reason why the two methods appear to disagree is that the correction for attenuation formula produces an estimate of the minimum single-item reliability. Thus, the “true” reliability could be, and probably is, higher than the estimated minimum. Because of this, both methods may be closer to each other than they first appear. In addition, factor analysis results in a point estimate of reliability, which means that it will be higher than a minimum estimate.

A third reason for the differences between the two methods is that the halo effect in the student ratings (to the extent that it is present in them) affects the two estimating procedures in different and disproportionate ways. This difference means that the factor analysis method may produce overestimates of single-item reliability if there is a halo effect in the student ratings.

An inflated estimate using factor analysis is possible because the communality estimates are dependent on the inter-item correlations. To the extent that there might be halo effects in the ratings, it would increase the consistency across all 10 items and, thus, inflate the communality estimate. Although halo effect can inflate both methods of estimating single-item reliability, it is likely to inflate the estimate based on factor analysis much more so than the estimate based on the correction for attenuation formula, as explained below.

To the extent that halo effect might be present, it would inflate the correlation between the overall single-item measure and the scale of the nine facets. It would also inflate the reliability of the nine-item scale of effectiveness facet items. Single-item reliability estimates based on the correction for attenuation formula rely on the correlation between the scale and single-item measure and the reliability of the scale. Thus, increasing the magnitude of the correlation between the scale and the single-item inflates this estimate. However, increasing the reliability of the nine-item scale deflates the estimate of single-item reliability. Because these two effects work in opposite directions, they tend to offset each other. Thus, the possible influence of inflation due to halo effect is fairly small—or nonexistent—when using the correction for attenuation formula. However, when factor analysis is used, halo effect (if present) causes an overestimate of reliability.

By way of comparison to the above estimates of single-item reliability, Arvey et al. (1992) reported single-item reliabilities for eight items that averaged to .66 using only the factor analysis method. The most likely reason why their estimates are lower than
those reported here is that their single-item measures are more diverse than the nine teaching effectiveness facets that can be represented in one factor. In their research, which concerned physical ability measurement, their eight single-item estimates concerned grip strength, dummy wrestling, dummy drag, a 100-yard dash, an obstacle course, sit-ups, bench dips, and a 1-mile run. Some of these concerned endurance, whereas others concerned strength, a distinction supported by confirmatory factor analysis (Arvey et al., 1992, p. 1002). The diversity among their items (two underlying factors) compared to the present study (one underlying factor) was reflected in much lower interitem correlations and, thus, lower reliability estimates using the factor analysis method. Furthermore, their single-item measures were based on physical tests rather than ratings by one person of another. Thus, they were unlikely to have halo rating error that would inflate the single-item reliability estimates.

**Differences Between Levels of Analysis**

Although the overall average of reliability estimates for the two methods and the two levels of analysis is .76, we believe that the two levels of analysis should be considered separately. The average of the two methods for individual-level data (.70) is lower than that for the group-level data (.82) and quite similar to the estimates for single-item measures of job satisfaction previously reported by Wanous et al. (1997) that were based on individual-level data. For the study of student ratings of college classroom teaching, however, the appropriate level of analysis is the group level, that is, intact classes of students whose overall mean represents their collective perception of classroom teaching effectiveness. Other situations similar to the present one, such as when survey data or effectiveness ratings are aggregated at the team or department level in organizations, should be guided by the group-level estimates found here.

Our initial belief was that the present research would yield higher reliability estimates because the data are from one organization, only one measure is used, and a more homogeneous group of people was involved, as compared to the Wanous et al. (1997) meta-analysis. We believed that having fewer sources of between-studies variance would produce a higher estimate. This belief seemed to be supported until we recalculated the estimates on the individual-level data at the request of the editor. Even though we all agreed that the individual-level analysis was inappropriate for this particular situation, the comparison made sense because of the relevance for other research domains. As can be seen from the results, the difference in levels of analysis appears to be the reason why higher estimates were found for the group-level data and not what we originally thought.

Using aggregated data does remove a source of variance, that is, differences among individuals. As has been persuasively argued in the research literature on student evaluations of teaching (Cranton & Smith, 1990), using group means for assessing classroom teaching effectiveness is appropriate because the purpose of student ratings is to differentiate among many different instructors. Using a class mean for student ratings tends to minimize the effects of extremely high ratings made by some individuals in the class that might be based on their own personal reaction to the instructor. Extremely low ratings can be made by students who disliked the course material, the instructor, the classroom, the time of day the class was held, and so on, even though none of these factors represents the classroom teaching effectiveness of the instructor. Using overall class mean ratings limits the effects of such extreme ratings that may be
extraneous to the actual effectiveness of the individual instructor. (This is also why we eliminated all classes from our analyses that had fewer than five students.)

**Conclusion About These Estimates**

As can be readily seen by now, it is difficult to estimate the reliability of the single-item measure with exacting precision. For the group-level data (that is relevant for this context), our best guess is that the minimum estimate of .70 may be too low because it is based on the most conservative assumption possible: a construct correlation of 1.00 between the single-item and the scale of nine teaching effectiveness facets. Similarly, the estimate of .94 from factor analysis might be too high. Based on the discussion about the assumptions and limitations of each method for estimating reliability and the fact that the average of the two methods is .82, we conclude that a reasonable single-item reliability estimate for group-level data is at least .80.

For the individual-level data, we conclude that a minimum estimate of .70 for the reliability of a single-item measure is reasonable. This is the same conclusion reached by Wanous et al. (1997) for a single-item measure of overall job satisfaction. Although the research domains studied are different (job satisfaction vs. teaching effectiveness), the conclusions for individual-level data are the same.

**Other Considerations: Research Context**

This is a study of student ratings of college instructors, which might be considered as a type of subordinate appraisal. Thus, it makes sense to address the similarities and differences between this context and that typically found in business. First, both are similar in that one person rates the behavior of another person. Second, the identity of the rater is anonymous. Third, these ratings are summed up and averaged before being reported to the course instructor or to the supervisor, because aggregating data is an additional way to protect the anonymity of subordinates. Fourth, both can be used for evaluative and/or developmental purposes, depending on the particular organization.

There are some differences between student ratings of classroom instruction and subordinate ratings of the supervisor. First, in business situations, most people have just one supervisor, and there is not too much turnover among supervisors—at least in the short run. In contrast, students typically take three to four courses per quarter at this university (there are three quarters in the academic year). As a result, students rate several different instructors at one time and do this three times a year over the 4 to 5 years it takes to graduate. Thus, the typical undergraduate student will rate 40 to 50 instructors during their years at this university. Graduate and professional students take fewer courses but still complete quite a few ratings. It is extremely unlikely that a person in business would have that many different supervisors over a similar period of time. Furthermore, a typical businessperson would only rate his or her supervisor on a yearly basis, so there would be five ratings over a 5-year period versus the 40 to 50 instructor ratings a student would produce over a similar period.

A second difference is that students have the opportunity to observe most, if not all, of the instructor’s classroom teaching effectiveness, unless they frequently “cut class.” To be sure, students are not in a position to rate how much effort an instructor puts into class preparation, because they only can observe what transpires in the classroom.
itself. In contrast, many subordinates in business may see only part of their supervi-
sor’s job performance. For example, they may not see their supervisor’s peer relation-
ships or their supervisor’s relationship with his or her own supervisor. In some situa-
tions, a supervisor’s direct reports may be scattered geographically, as is typical in
sales departments, further limiting the opportunity to observe one’s own supervisor.

Students are asked to rate classroom teaching effectiveness, and they are generally
in a good position to do just that. They can observe all, or most, of the behavior to be
rated, and they can fine-tune their ratings because they have the opportunity to observe
a large number of instructors. In addition, the “contrast effect” from observing many
different instructors can be beneficial in the sense that it provides actual role models of
the differences between good versus poor instruction. In contrast, the typical business-
person may not experience nearly as many supervisors even in an entire working
career. Those who start out in business and whose first supervisor is ineffective may
not realize what is being missed until they experience a much more effective person
later in their career. In contrast, students see many role models, which helps them
develop their own ideas of what constitutes effective classroom instruction.

It might seem surprising that no moderators were found for the correlation between
the two measures of overall teaching effectiveness (i.e., the single item vs. the sum of
nine facets) because 35% of the between-class variance was explained by the two arti-
facts accounted for in the bare bones meta-analysis (sampling error and measurement
unreliability). This leaves 65% of the between-class variance unexplained. However,
there may have been unmeasured moderator variables that could account for this unex-
plained between-classes variance. Examples of such variables could include instructor
teaching styles, student learning styles, and instructor grading standards. Although
plausible, data of this type could not be obtained to examine these possibilities.

Limitations

The present research effort does have several possible limitations. First, the
generalizability of results may be limited because all data were collected at one univer-
sity. That said, there is no reason to suppose that this university is substantially differ-
ent from other large research universities located throughout the country. Furthermore,
the sample size studied here is extremely large. In fact, the sample size for this study is
substantially greater than that found in almost all previous research in industrial/orga-
nizational psychology or organizational behavior.

A second possible limitation is that the student ratings of teaching effectiveness
may have been biased by halo effect. If true, the presence of halo effect would affect
the estimating method based on factor analysis disproportionally more so than that
based on the correction for attenuation formula. It is not possible, however, to estimate
the degree to which halo is, or is not, present in these data. Furthermore, it is not possi-
ble to estimate the extent to which halo, if present, represents error variance or true
variance. Examining Table 1 shows that there is variance among the various facets of
teaching effectiveness in terms of mean values and in terms of the correlation of each
facet with the overall item (from .70 to .93). This would seem to indicate that there may
be less rather than more halo error.

Instructors at our university have noted from time to time that ratings of “inter-
ested” and “helping,” for example, vary considerably when a large (N = 100 to 400),
required undergraduate course is being taught versus a small \((N = 25\text{ to } 40)\) elective course for one’s own majors. These differences, albeit anecdotal, would seem to reflect true differences in these two very different teaching/learning environments.

A third possible limitation is that data from students do not generalize beyond colleges and universities. There are several responses to this possible limitation. First, the number of higher education institutions represents a rather large sector of the American economy. Even if these findings were strictly limited to this sector alone, they would still be quite important due to the sheer size of the college and university sector in this country. Second, students are quite experienced in using the SEI form because it is a university requirement for every course. Students may be young and inexperienced in some ways, but they are not inexperienced with the SEI.

Finally, the single-item measure of overall teaching effectiveness was compared to only one other measure of overall teaching effectiveness, that is, a summed scale of nine facets. There was no opportunity to compare this single-item measure to a scale based on items that have an overall teaching effectiveness focus instead of a facet focus, and there was no opportunity to compare it to a different set of teaching facet items. We were limited by the university’s choice to use these particular nine facets.

The possible difference between a summed scale of facet items versus a summed scale of globally focused items was, however, assessed in the previous work on single-item reliability (Wanous et al., 1997). In that study, the results for scales of job facet satisfaction versus scales of overall-focused satisfaction items were almost exactly the same. The average correlation between a single-item measure of overall job satisfaction and a scale of specific job satisfaction facets was .64. By way of comparison, it was .67 when the single item was correlated with a scale of globally focused items (Wanous et al., 1997, p. 249). This appears to be a trivial difference, so the only question is whether it would be likely to be replicated here as well.

There is a consensus that the validity of an operational measure is the most critical element of its psychometric properties. The present study did not address validity directly. Psychological measures cannot be valid unless they are first reliable. Unfortunately, too many people have assumed two things about single-item measures: (a) that the reliability of single-item measures cannot be estimated, and (b) that the reliability would be unacceptably low, if it could be estimated.

Previous research on job satisfaction by (Wanous et al., 1997; Wanous & Reichers, 1996) showed how to estimate single-item reliability and that its minimum level was reasonable for overall job satisfaction. In the present study, we extend this finding to a new construct. Furthermore, we do so with data that most likely provide a better estimate of minimum-level, single-item reliability because four sources of between-study variance are controlled here. Our conclusion of \(r = .80\) as the minimum level of single-item reliability for group-level data was based on two different methods, while taking into account the strengths and weaknesses of each method. Our estimate of \(r = .70\) as the minimum level of single-item reliability for individual-level data supports previous research.

References


John P. Wanous (Ph.D., Yale University, administrative science, 1972) is a professor of management and human resources in the Fisher College of Business and a professor of psychology in the College of Social and Behavioral Sciences at The Ohio State University, positions he has held since 1983. He was previously on the faculties of NYU and Michigan State University. He is best known for his extensive work on organizational entry. He is the editor of the Addison-Wesley (now Prentice Hall) book series titled Managing Human Resources, a position he has held since 1977.

Michael J. Hudy (Ph.D., The Ohio State University, psychology, 1997) is a senior consultant with SHL USA, Inc. at its Cleveland Office. The data for this study came from his doctoral dissertation.
A Community of Voices: Using Allegory as an Interpretive Device in Action Research on Organizational Change

JOSEPH W. GRUBBS
Grand Valley State University

Organizational change involves patterns of engagement across organizational lines and thus carries important cultural implications for participants and groups within communities of organizations. This article proposes using allegory as an interpretive device in the study of organizational change. Allegory allows organizational research to appreciate the way “characters” introduce distinct qualities into the change experience. Based on an action research program with public organizations in the state of Delaware, the article reveals the promise of allegory both for explicating theory and informing reflexive practice.

Allegory transforms experience into a concept and a concept into an image, but so that the concept remains always defined and expressible by the image.

—Goethe

Change has become a central theme in the study of organizations. During the past decade, scores of articles have been dedicated to exploring the process and management of change (Kahn, 1993; Kim, 1993; Newman & Nollen, 1998; Van de Ven & Poole, 1995), and an entire industry has emerged around the practice of change management (Worren, Ruddle, & Moore, 1999). Although the interest in change among organizational scholars continues to rise, our understanding of the concept remains limited by several flawed assumptions relating to change and to the nature of organizations.

One of the most concerning assumptions, particularly in today’s world, leads us to view change as a process that occurs one organization at a time, with our inquiry dedicated to observing change within individual organizations (for examples, see Kotter, 1996; Nadler, 1998; Newman & Nollen, 1998). Such an image fails to reflect the increasingly complex nature of organizations in the postindustrial era. As organizations from all sectors continue to forge joint ventures, strategic alliances, and other collaborative relationships, we must think of change as involving and affecting entire organizational communities (Bardach, 1998; Marks & Mirvis, 1998; O’Toole, 1997; Provan & Milward, 1995; Spekman & Isabella, 2000).
A second flawed assumption constricts our inquiry to the mostly functional, rational aspects of change. Consequently, even when we take into account the dynamic nature of organizational alliances, our focus often remains on issues of organizational design, transaction cost, or the resource dependency underlying the relationships (Alter & Hage, 1993; Bluedorn, Johnson, Cartwright, & Barringer, 1994; Elg & Johansson, 1997; Oliver, 1990; Pfeffer & Salancik, 1978; Robins, 1987; Thorelli, 1986). We fail to see change as a symbolic process, one that carries important cultural implications for participants and their organizations (Grubbs & Denhardt, 1999; Lawson & Ventriss, 1992; Trice & Beyer, 1984).

In this article, I suggest that organizational scholars should adopt a richer, more meaningful image of change by exploring the symbolic aspects (Alvesson, 1995; Jones, 1996; Smircich, 1983, 1985; Trice & Beyer, 1984), where participants act according to diverse cultural attributes within their respective organizational communities (Czarniawska, 1997a; Rosaldo, 1989; Wilkof, Brown, & Selsky, 1995). By understanding change in its symbolic context, it is possible to identify the mostly tacit barriers to change and pattern approaches that reflect culturally influenced interpretations of change processes.

On the other hand, such a goal is limited by similar flawed assumptions as those affecting research into organizational change—namely, that our interpretive inquiry most often is conducted within individual groups. Although we may participate in important research into the symbolic aspects of organizations, we often do so one organization at a time. Our attempt to appreciate cultural influences across multiple groups usually involves piecing together findings from analyses of the respective organizations. The question is, how can we adopt methods of organizational research to take account of the complex nature of change while appreciating the cultural significance of change across organizational communities?

This article discusses my use of the literary form, allegory, as a narrative device to examine the symbolic aspects of change across an alliance of public sector organizations in the state of Delaware. My role was as an action researcher, and not as an ethnographer, and my purpose was to understand the cultural factors affecting the change and then to use this insight to inform my intervention into the change process. At the heart of my action research program was the need to look beyond the more rational, structural issues and to unveil the tacit barriers, which emanated from the belief systems of participating organizations (Argyris, 1993; Argyris & Schon, 1978, 1985; Grubbs & Denhardt, 1999). To do this, I used qualitative methods to explore the cultural attributes of participating organizations and the way these attributes influenced the shared-change experience. Allegory, as an interpretive device, provided a lens that allowed me to more effectively view the cultural factors affecting the change process.

From the standpoint of literary theory, allegory is an extended metaphor in which authors employ narrative devices to convey a more symbolic meaning than is otherwise apparent in the text (Madsen, 1996; Ortony, 1993). In the context of my research, allegory was used at various points in the action research program to analyze findings from my qualitative inquiry and to present these findings, along with my interpretations, to the other participants. Allegory first offered a way for me to make sense of the various themes from the change, in particular the images of the change as expressed by participants based on their respective organizational belief systems. For example, the more traditional agencies, whose principal value tended to be efficiency, spoke with a "voice" that characterized the change as a way to increase productivity and cut operat-
ing cost; agencies that were more concerned with effectiveness expressed themselves in a voice oriented more toward enhancing the quality of service.

As I examined participant narratives during the creation of the research text, I framed my interpretation around these voices to see how they reflected the cultural influences underlying the change experience. Allegory as an interpretive device provided a forum in which the metaphorical characters could interact and through which I could appreciate the symbolic significance of this interaction. During this phase, I was able to place the characters into a fictional plot and to see the way organizational assumptions and value systems came into contact during the story of change.

Allegory also was used during the creation of my public text, the actual narrative I revealed to participants and other readers. I presented the allegory first by itself, allowing the reader to experience the symbolic narrative in an abstracted form, removed from the local context. Next, I provided overlays of my interpretive comments and of actual statements from the participants, thus enabling the reader to see my interpretations of the mostly symbolic aspects of the change and to recognize how these aspects influenced the change process. Through this public text, I was able to stimulate a dialogue relating to the tacit barriers to change and to facilitate consensus building across the organizational community on more meaningful, culturally sensitive patterns of change.

On Allegory

Goethe’s quote in the epigraph reveals the symbolic power of allegory, in particular its capacity to transform human experience first into an abstract notion and then into an enduring image. The etymology of the word stems from the Greek root allos or “other” and agoria or “speaking,” with the term other-speaking referring to the presence of a level of meaning below the surface of the text (Leeming & Drowne, 1996; also Clifford, 1986, p. 99). In its classical form, allegorical literature can be traced to Greco-Roman and Judaic origins, and it features the use of literary devices to communicate moral, philosophical, or other messages beneath a veil of abstraction (see also Whitman, 1987).

Take as an example two allegorical narratives from Western literature: John Bunyan’s The Pilgrim’s Progress (1678/1987) and Nathaniel Hawthorne’s Young Goodman Brown (1835/1967). On the surface, the works share common characteristics. Both depict their protagonists on a journey toward some mystical experience, with symbolic devices used to reveal various religious themes. Yet, the “other meaning” of the allegory tells a different tale. Bunyan’s is the story of Christian, a personification of the model Puritan whose path symbolizes the trials of the righteous in the earthly domain. Hawthorne’s, on the other hand, offers a critique of Puritan society, a tale of corruption, spiritual guilt, and moral hypocrisy. Although similar in form, the narratives yield very different images of early Anglo-American religious practice (for discussion, see Grubbs & Denhardt, 1999).

In the study of culture and organizations, several authors have employed allegory as a way of discussing diverse levels of meaning. The literary form, in this regard, most often has been adopted as an interpretive device—a framework to help make sense of the different frames of reference underlying the respective fields. For example, Clifford (1986) suggested that the entire practice of ethnography could be viewed as allegorical in nature. He wrote,
Ethnographic texts are inescapably allegorical . . . [and] the very activity of ethnographic writing—seen as inscription or textualization—enacts a redemptive Western allegory. . . . Allegory (more strongly than “interpretation”) calls to mind the poetic, traditional, cosmological nature of such writing processes. (pp. 99-100)

Citing Shostak’s classic study, Clifford traced the diverse allegorical voices in the ethnographic narrative. These voices, Clifford concluded, reflect the different meanings that occur, both for the ethnographer and the reader, and have significant implications for the knowledge that follows from the ethnographic experience.

For Clifford (1986), the principal concern is how the meaning from ethnographic accounts transcends the “simple” incidents in their local context and, with the multiple layers of interaction between author, reader, and text, becomes part of a broader narrative of human experience. The cultural norms and lifeways of the !Kung described by Shostak, in particular those of the woman, Nisa, go beyond their time and space significance—that is, a point in time in the Kalahari Desert. Due to the process of ethnographic research, as it reveals the many voices that appear in the text, these elements of culture become transformed to have what Clifford called “transcendent meaning” (p. 99). He wrote,

[Allegory] draws attention to aspects of cultural description that have until recently been minimized. A recognition of allegory emphasizes the fact that realistic portraits, to the extent that they are “convincing” or “rich,” are extended metaphors, patterns of associations that point to coherent (theoretical, esthetic, moral) additional meanings. (p. 100)

In organization theory, Frost, Moore, Louis, Lundberg, and Martin (1985) used allegory to present the findings from their meta-analysis of the study of organizational culture. The public text featured a fictional narrative in which different types of birds served as symbols for the way scholars interpret culture in human organizations. The narrative began,

Several spirit birds were flying together, returning from an extensive trip to investigate the cultures of human organizations. The birds had been flying for a long time and, as they were weary, they stopped to rest on a large rock. In the group were a raven, an owl, a wren, an eagle, a stork, a pelican, and a phoenix. In popular folklore, each of these birds is associated with distinct qualities . . . Each bird was attempting to describe cultural experiences it had had during its travels, for each had found the cultural aspects of human organization particularly fascinating. (pp. 13-14)

The symbolic language of allegory enabled the authors to reveal how different frames of reference may contribute to vastly different understandings of culture, in their case based on an array of papers presented at a conference. They wrote, “Like the individual birds in the [allegorical narrative], each conference participant brought along a different perspective on organizational culture” (p. 24).

As with the allegory of ethnography described by Clifford (1986), the different voices present in Frost et al.’s (1985) symbolic narrative of organizational culture take on meaning at several levels. In one sense, the figures represent the different viewpoints from the particular conference, and as such they reflect the various images of culture in the field of organization theory. This ties the narrative to the local context of the participants. However, the true meaning comes out as these characters, reflections
of the different viewpoints, interact in a symbolic discourse. This is the other meaning of the allegory, or the “transcendent meaning,” to use Clifford’s phrase, in which the significance of the engagement extends beyond the local context. The narrative is not a history of the conference, nor is it a rational summary of the conference papers. Rather, it is an extended metaphor that takes on meaning at a variety of levels, and it is this broader meaning that allegory enables the authors to capture and present to the reader.

Weick (1996) also adopted allegory for the purpose of engaging in a meta-analysis of existing scholarship, in this case on the 40th anniversary of the publication of Administrative Science Quarterly. He used the Mann Gulch and South Canyon wildfire tragedies, in which firefighters refused to drop their “heavy tools” and consequently lost their lives, as a reflection of the traditionalism in organization theory. He wrote,

Dropping one’s tools is a proxy for unlearning, for adaptation, for flexibility, in short, for many of the dramas that engage organizational scholars. It is the very unwillingness of people to drop their tools that turns some of these dramas into tragedies. . . . Social scientists refuse to drop their paradigms, parables, and propositions when their own personal survival is threatened. To drop one’s tools, then, is an allegory for all seasons that is capable of connecting the past with the present. (pp. 301-302)

For Weick, this symbolizes the tendency of organization theorists to retain their conceptual baggage and, as a result, remain stagnant in their studies of human organization.

My use of allegory differs in scope from the other adaptations in ethnography and organization theory, but the principles remain the same. The purpose of borrowing the literary form is to use it as an interpretive device for exploring processes of change across communities of organizations. As such, allegory can be understood as an extended metaphor in which characters or objects in the narrative represent unique cultural attributes from their respective organizational communities (on the use of complex metaphor, see Czarniawska, 1997b). By cultural attributes, I am speaking of the social influences of human action, what Geertz (1973) referred to as “plans, recipes, rules, instructions” (p. 44). These attributes allow group members to make sense of their surroundings, to adapt, integrate, and respond to various factors in their social settings. Over time, such attributes become shared within the group and manifested in the beliefs and actions of individual members (Ingersoll & Adams, 1992; Schein, 1997).

Despite the inherent cultural diversity, organizations that have undergone a similar formative process or that share comparable characteristics in their value systems may exhibit some of the same cultural attributes. Human actors from these groups may come together with representatives from other groups, based on the shared assumptions and beliefs underlying their otherwise distinct organizational cultures. The cultural attributes that become shared across organizations in networks and that emerge around issues of organizational change serve as the building blocks of the allegory. The literary form offered a way for me to understand these influences and their implications for organizations engaged in a shared-change experience. Specifically, allegory was used as an interpretive device that enabled me to unravel the oral and textual narratives from the participants and to appreciate the various cultural factors from the community of organizations.
Case Study: Allegory and Change in the State of Delaware

To explicate the notion of allegory as an interpretive device, it may be helpful to consider a case study based on my involvement in a change process in the state of Delaware. From 1996 to 1998, I served as a consultant to the state government to facilitate a change initiative involving several of the agencies responsible for human and social service delivery. My role, along with several other colleagues, was to conduct research and analysis on various aspects of the change and to inform the process through research reports, training seminars, and ongoing process and management consultation.

The state of Delaware’s change initiative began in 1993, when Governor Tom Carper issued a directive challenging the state government’s human and social service agencies to develop an integrated system of service delivery (Carper, 1993; see discussion in Grubbs & Denhardt, 1999). Carper’s executive order brought together cabinet-level secretaries from seven of the state government’s lead service organizations, a group that became known as the Family Services Cabinet Council. With backing from executive leadership, most of the functional details of the change process were carried out by executive staff from the Cabinet Council agencies, a group known as the Services Integration Working Group.

My action research consisted of the use of qualitative methods, with a document search, semistructured interviews, and participant observation as my primary strategies for gathering information. First, I conducted a document search to collect historical resources concerning the state of Delaware’s service integration agenda. This helped me to appreciate the organizational community in its broader social and political context. Documents reviewed included letters, e-mail messages, formal memos, and other textual narratives, which were either associated with the change process or with other activities related to the effort. In addition to their substantive relevance, these documents were examined as cultural artifacts, symbols reflecting the deeper significance of the transformation within the Delaware community.

Second, I conducted semistructured interviews with executive staff from the Services Integration Working Group. I determined that this would be a key level of analysis because the group consisted of the primary architects of the change initiative and represented a microcosm for the formation of culture within an organizational community. Despite ties to their respective organizations, working group members for several years had engaged on this particular change process, as well as other policy and administrative issues. I found that such engagement had enabled them to establish a sense of trust and shared communication, important characteristics for the formation of an organizational community (Adams & Ingersoll, 1990; Schein, 1997).

Third, a similar group of semistructured interviews was conducted with cabinet secretaries who make up the Family Services Cabinet Council. These individuals served as a source of executive leadership for the network in the overall change experience and therefore offered a unique perspective on leading change and facilitating collaboration within the emerging social service network. In a broader sense, interviews with these individuals provided valuable insight into the nature and role of leadership in the state of Delaware.

My goal during the information collection phase involved identifying the key sources of oral and textual narratives, then gathering from these sources narratives
relating to the change process. Of interest also were my field notes, which contained important information and reflections from my engagement with the research participants. The narratives, both from the participants and myself, were then catalogued, analyzed, and coded through a process of selection that enabled me to highlight diverse images of the shared change.

During the research process, I faced several important ethical challenges. First, although wanting to maintain the confidentiality of the research participants—for a number of reasons, including protecting them from any negative consequences resulting from their participation—it was my responsibility as an action researcher to ensure accuracy in the research findings. My use of tape recordings and field notes helped in this regard, as I was able to compare my analysis with the interview transcripts and to consider my “readings” in light of the tone and context of information gathering events (Altheide & Johnson, 1998).

Second, as an action researcher, I was obligated to conduct my inquiry on the basis of informed consent on the part of the participants. This was obtained either using a formal document that explained the confidential and voluntary nature of the research, which the participant was asked to sign, or through a verbal statement at the beginning of the interview, to which the participant would acknowledge in the taped record. (For a discussion of ethical issues, see Christians, 2000.)

The use of allegory as an interpretive device occurred during several phases of my research program: the interpretation phase, the development of the research text, and the final preparation of the public text from my inquiry (Denzin & Lincoln, 1998; also Van Maanen, 1988). As Czarniawska (1997b) suggested, I employed the literary form as an extended, complex metaphor to understand very complex phenomena surrounding processes of change and to communicate this understanding to the various readers. The purpose of allegory as an interpretive device thus was twofold, first to interpret the themes emanating from the change experience, and second to initiate a dialogue among participants that allowed for a more effective facilitation of the change.

Allegory was first used as a narrative device during the interpretation phase of the research, that is, as a tool to help me make sense of the diverse narratives emerging from the change process. As I reviewed narratives from the participants and from myself as the researcher, I began to notice distinct themes. These themes reflected different images of the change as held by the participants. Moreover, as I delved deeper into the oral and textual narratives, I began to notice that the themes were not random but reflected the belief systems underlying participating organizations.

Central administrative agencies, for example, such as the Budget Office, and the more politically oriented agencies, such as the Governor’s Office and the legislative offices, viewed the change as a way of reducing service redundancy and increasing fiscal and administrative efficiency. On the other hand, service-oriented agencies, such as the Department of Services for Children, Youth and Families, saw the service integration initiative as a way to more effectively provide support for children, families, and communities. Other voices also came forward, such as those focused on the “how” of service integration, with an emphasis on training, technology, and facilitation.

By adopting the narrative device of allegory, I transformed these themes into “characters.” The characters were developed according to the values being communicated by the participants. In other words, the characters served as symbols of the cultural attributes of participating organizations and, in turn, reflected the way these attributes became manifested during the change process. However, the characters should not be
viewed as metaphors of a single organization or participant; rather, they reflected the cultural attributes shared across organizations in the community, personifications of the cultural attributes of organizations with similar value systems. Although beliefs within each respective group varied and each group had its own distinct cultural system, important characteristics from these meaning systems were shared between the groups.

Table 1 provides a list of the characters from my interpretation of the change experience and the cultural attributes they symbolized. For instance, charity personified the more service-oriented agencies within the state of Delaware, agencies whose primary values centered on serving children and families. Members of these organizations interpreted the change experience as a way of enhancing the quality and accessibility of state services for families. Thrift, on the other hand, represented those agencies whose primary values focused on fiscal efficiency. Agencies with cultural attributes portrayed by thrift tended to view the change as a way of “doing more with less.”

The allegorical figures became lenses for me during the interpretive phase of my research, devices that enabled me to appreciate the ways in which cultural attributes from diverse organizations affected the change process. When I interviewed participants or combed through documentary texts, I heard the voices of these characters speaking to me (figuratively, of course!), communicating important insights concerning themselves and their organizational value systems. As the interpretive phase of my research unfolded, I was able to engage in a dialogue with these characters, thereby gaining a richer appreciation of the frames of reference underlying the organizational drama.

My use of allegorical characters was not some reductionist attempt to set parameters on, or to generalize, the diverse meaning systems within participating groups. The goal was not to construct organizational or cultural “types.” Even if this were possible, which from my standpoint it is not, such an approach would fail to take account of the unique system of beliefs within each group (Alvesson, 1995; Frost et al., 1985; Sackmann, 1992, 1997; Schein, 1983, 1997). Instead, the use of allegory enabled me to interpret the ways in which attributes from these belief systems emerged around, and in many ways influenced, the change experience. The attributes were shared by multiple organizations, due to common traits in the groups’ meaning systems, and reflected the way members of these groups made sense of the change.

The second use of allegory occurred during the preparation of the research text, or the interpretative narrative I prepared to make sense of the symbolic meaning underlying the change process (Denzin & Lincoln, 1998). As I engaged in my interpretation of the qualitative material, I saw the characters begin to form as symbols of the various sets of cultural attributes. Besides being important as symbols, however, the characters took on a deeper significance through their interaction in the shared-change experience. In this regard, the allegorical narrative I created as the researcher, my research text, became the forum in which the characters could interact.

My development of the research text actually became a process of creative writing. In fact, the research text itself was a fictional narrative. Before endeavoring in this process, I read allegories from several authors to see the way they introduced characters, designed plots, presented action, and other literary techniques. Thanks to this review of allegorical literature, I began to see the characters in a new light. They ceased being simply metaphors for the various frames of reference, but in a figurative sense, they became living and breathing creatures. My desire was to get to know these characters,
appreciate them at many levels, and enlist their support in my understanding of the way the different frames of reference in the organizational community affected the change process.

In the research text, I placed the voices into a dialogue—one that, although removed from the local context in the state of Delaware, provided a glimpse into the Delaware change. The allegorical figures were allowed to roam freely in their fictional universe and to engage with each other based on their respective frames of reference. In doing so, they yielded valuable insight into the factors affecting the change process. Allegory, as an interpretive device, provided the narrative stage for me to experience and then to understand the “other meaning” of the engagement between the various systems of belief.

The research text also provided the forum for me to construct a variety of scenes to determine how the service-oriented character, Charity, would interact with the more efficiency-minded Thrift. Would the two characters form a consensus and join together in the journey? Or, would they remain distant from the other, making attempts
to effect the change based on their own frame of reference? And what of the other characters? What part would they play in the change? The allegorical form enabled me to appreciate the characters as personifications of their respective value systems but also to recognize the way they added to a deeper meaning of the change experience. Through the allegory, I captured the pervasive impact of participants’ meaning systems on the change and to understand the implications of this impact for the organizational community. Moreover, I was better able to see my influence as the researcher as I become a character in the allegorical narrative.

The third use of allegory occurred during the creation of the public text, or the narrative that I presented to the other participants and to the scholarly community. My purpose for using allegory in this phase was to highlight for discussion the mostly symbolic aspects of the change but to do so without raising defensive routines among the participants. Of course, this required careful consideration of my relationship with the reader. As Rabinow and Sullivan wrote, texts are “plurivocal, open to several readings and to several constructions” (cited in Riessman, 1993, p. 14), and I wanted to use the forum of the public text to form a bond with those who may come into contact with the public narrative. The readers would serve first as filters, interpreting the story based on their sociocultural context, then as conduits between the text and others around them (Clifford & Marcus, 1986; Riessman, 1993).

To present the allegory in the public text, I adapted a framework similar to that used in the learning history, a process developed by scholars in the Society for Organizational Learning (Roth, 1996; Roth & Kleiner, 1995) and similar to Argyris’s (1993) approach to providing feedback in an action research program. Specifically, I structured the presentation in three ways: first, the allegory by itself as a prose narrative; second, with an overlay of my interpretive comments; and third, with actual excerpts from the participants in the Delaware change. In this way, the allegory was offered first as an abstraction from the individuals and events surrounding the shared-change process, then along side the overlays, which revealed the story’s other meaning.

Tables 2 and 3 show selections from the allegory in the left column, with my commentary and the actual statements from participants on the right. The fictional narrative allowed readers to experience the symbolic aspects of the change, whereas the overlay of my interpretive comments enabled them to appreciate how these symbolic aspects affected the change process. The selections in Table 2 concentrate primarily on contextual issues. My commentary in this regard represented an overall reading of the change experience, with the primary intent of communicating my interpretations of the purpose and scope of the change. In this way, participants were encouraged to engage in reflexive practice—that is, to look at how their own interpretations of actions and events contributed to the way these events were constructed within the organizational community.

The other meaning of the allegory becomes more apparent in Table 3, as I introduce the allegorical characters. Here, the reader is able to see the characters as symbols, reflections of the various cultural attributes, and the way the characters interact, based on my interpretation of the attributes that they personified. As discussed previously, such elements of the allegorical narrative were derived through analysis of the communication patterns that I experienced during the research process. My approach involved using the overlays of transcripts from the semistructured interviews to further illustrate the connection between what was said and the value system that contributed to the narrative’s meaning. The intent, to reveal the deeper significance of language, artifacts
and other organizational symbols, and the way these cultural aspects were reflective of the belief systems within participant groups.

The actual unveiling of the allegory in this particular case was indirect; that is, the public text was simply made available to participants for general review and discussion. We did not hold a separate meeting or work session to share our interpretations. This proved to be a major limitation as it failed to generate the type of dialogue necessary to affect the change. The benefit of the allegory, however, even with this limitation is that it provided me with greater insight into the cultural factors affecting the change. In turn, such insight enhanced my effectiveness as a facilitator of the change process. I was able to identify the tacit barriers to change and to inform the level of action among participants necessary to overcome these barriers.

Table 2

Interpretation of Network Context

| Charity awoke from a horrible dream: She had seen a family wandering in the wilderness, weak and without food or shelter. After hours of stumbling across the barren landscape, the family came upon a house. It was not an assuming house, its sides and roof weathered, but it seemed larger than it actually was because of its imposing front door. Cast iron and a full 2-feet thick, the door was like an entranceway one might have found on a medieval castle, or a walled city, impene- trabrable. The father of the family reached with his tired hand and knocked as hard as he could. But given the weight and breadth of the door, his knock barely rose above the tempest that scoured the wilderness behind them. Again he knocked, but knew that certainly no one would hear. Finally, when the family had turned to leave, a sound of latches being unfastened came from the other side. The door cracked open, and a face peered through. It was not an unfriendly face, but it certainly did not show the warmth the family had hoped for; it was not a face that promised comfort from the storm. “May I help you?” asked the face, for the family could see nothing more. “My family has been ravaged by the storm,” the father said, trying to keep his voice from trembling. “We have no food or shelter. Our children are in need. Can you help us?” “We do not offer food or shelter here,” said the face. “And there is no doctor to help your children. We have only water. For the other things, you must continue through the wilderness to the other houses. At each, you will find some of what you need.” |
| The current array of services, marked by a fragmented and complex system of delivery, as expressed in Governor Carper’s (1993) executive order and the Lochtenberg Commission (1993) report. The imposing nature of government service agencies, particularly those that remain bureaucratic in form. The fact that much of the labor required for accessing services falls on families as opposed to agencies offering services where they are needed most. Service delivery systems are not family friendly, instead retaining an agency-based focus. The categorical, issue-based form of service delivery, which forces families to search out the support they need by going from agency to agency. |

and other organizational symbols, and the way these cultural aspects were reflective of the belief systems within participant groups.

The actual unveiling of the allegory in this particular case was indirect; that is, the public text was simply made available to participants for general review and discussion. We did not hold a separate meeting or work session to share our interpretations. This proved to be a major limitation as it failed to generate the type of dialogue necessary to affect the change. The benefit of the allegory, however, even with this limitation is that it provided me with greater insight into the cultural factors affecting the change. In turn, such insight enhanced my effectiveness as a facilitator of the change process. I was able to identify the tacit barriers to change and to inform the level of action among participants necessary to overcome these barriers.
Assessing Allegory in Organizational Research

The principal strength of allegory as an approach to narrative analysis in organizational research is that, first, it enables participants to identify key attributes of the values underlying individual groups and then to understand how these attributes affect organizational members’ views on the shared-change process. Through a group’s patterns of communication, as well as its rituals and myths, insights can be gained concerning the organization’s belief system. Points of interest here include how group members share information about themselves, interpret actions and events, celebrate

Table 3
Interpretation of Participant Narratives

As Charity left her house, she came across Thrift, who had experienced a similar dream. Thrift, too, had seen that problems existed in the way people in the land of well-being supported families and recognized that something had to be done.

Charity: “Thrift, I have had a terrible dream. I saw a family struggling in the wilderness, trying to find the help they needed, until one by one they died in misery. Our many houses did nothing to help them.”

Thrift: “I had a similar dream. I saw our houses leave food, water, and other things outside for a family to partake on their own. And, some houses even offered some of the same resources. This must be changed.”

Charity: “We have to work together, to get our houses to share resources and offer food and shelter in ways that families don’t have to wander around through the wilderness. Maybe we can even give them what they need all in one place.”

Thrift: “I don’t know, Charity. I think we need to be more efficient in the way we offer food, shelter, and those other things. We can’t just keep all these houses around, or let families keep taking what they need. It just costs too much and the houses keep giving the same thing.”

Charity: “Oh, Thrift. I think we just have to be more supportive. It’s not the efficiency we should be concerned about as much as it is how effective we are in supporting families, and how simple we can make it for them.”

Thrift: “I still think it’s all that duplication, one house doing what some of the others are doing, then the fact that they just leave that precious food and water on the doorstep. I mean, families can just keep taking all they need.”

In the state of Delaware initiative, two distinct viewpoints emerged relating to the goal of service integration. These can be seen in the following statements:

Member of the Services Integration Working Group: “We need to do better than we’re doing, we need to integrate our services. We need to make it less complex for the client and family to navigate . . . taking the complexity and putting it behind the counter, so that the client can focus on their problems and not focus on mastering the system” (confidential conversation, September 18, 1997).

Participant in the state of Delaware’s deliberations: “It seems to me . . . you’re trying to maximize productive uses of your limited resources. And, to the extent that you engage in the duplication of services . . . you’re being wasteful and not providing the quantity of services you could optimally provide. That’s what’s been the problem” (confidential conversation, October 30, 1997).

Working group member: “Despite the fact that we have this marvelous system of state service centers, which is providing the government’s services for the most part . . . [state services] were complex. It was a challenge to manage those . . . it was all out there, but it was very hard for the client to navigate through them” (confidential conversation, September 18, 1997).
and punish behavior, and the tacit system of values that contribute to these manifest actions (Bell, 1997; Ingersoll & Adams, 1992; Schein, 1997; Smircich, 1983).

Second, allegory as an extended metaphor unveils the hidden meanings that emerge in the context of organizational alliances. As members of organizational communities, we are influenced in our interaction by factors in our respective meaning systems. Most often, these factors are not discussed openly, nor are they immediately recognized without a high degree of reflective practice. Analysis of these factors through the interpretive device of allegory provides organizational research a more suitable approach to understanding the “webs of significance” (Geertz, 1973, p. 5) we as humans construct to make sense of our social and organizational world (Chapple, 1941; Weick, 1995).

Third, the strength of allegory stems from its capacity to capture the “plurivocal” nature of organizational change, particularly as it occurs within communities of organizations. The multiple voices expressed in the allegory allow for a vast range of interpretations. For some, this may seem like a daunting proposition, as if the allegory was releasing the reader into the world with just enough interpretation to be dangerous. Such concerns, of course, have been expressed for interpretive social and organization theory generally (Burrell & Morgan, 1979; Geertz, 1973) and will no doubt be voiced toward an allegorical perspective. And yet it is this characteristic of the allegory as narrative analysis that I believe will allow the approach to find its way into future research, particularly action research and other forms of reflective inquiry.

Although allegory has great potential for revealing the attributes of culture within communities of organizations, one of its limitations is the superficial simplicity of the story. On one hand, the surface simplicity, by making the story accessible and clear, is exactly what facilitates the communication of the deeper meaning. On the other hand, the relative level of simplicity means that a limited number of characters and contexts can be included in the narrative. For the qualitative researcher, this means that decision rules must be developed to determine which themes or values will be highlighted in the allegory. In addition, in an attempt to distill the narrative interpretation to fit the requirements of an allegorical presentation, the richness of the narrative on which the allegory is founded may be lost (for a discussion, see Eco, 1990).

Likewise, just as with the first strength, the value of allegory as an extended metaphor has a concomitant requirement; that is, the researcher must be a competent writer of fiction for the allegory to successfully unveil the attributes of organizational symbolism. For those lacking experience in the preparation of fictional narratives, this presents an entirely new set of challenges. And beyond the creative element, researchers also must be able to balance the fictional side of the presentation with the analytical side so that the story is clearly grounded in and warranted by the narrative information. A concern here is that the researcher could become lost in the creative element, neglecting the underlying purpose.

It should be noted also that the concept of allegory proposed here should be distinguished from psychoanalytical approaches to organizational inquiry. For example, some have compared my allegorical characters to Jungian archetypes. Although I would never suggest excluding the cognitive dimension of cultural studies, my primary focus when interpreting the narratives lies less on the individual participant and more on the shared cultural attributes that emerge around the issue of the change. Cer-
tainly, the psychological type of each actor may be important at another level, but for the purpose of my inquiry, I remain more concerned with the cultural influences stemming from organizational meaning systems that affect the way human actors engage in the shared-change experience.

A distinction also should be made between allegory and previous efforts to employ literary forms in the presentation of organizational research, particularly Czarniawska-Joerges and Jacobsson’s (1995) use of the populist dramatic style, *commedia dell’arte*. In many ways, the work should be celebrated for its approach, which was based on literary interpretation and included an actual play with well-developed characters. My concern, however, is that by retaining a close connection between fiction and reality, the narrative may have threatened its own success. The veil of abstraction offered by allegory, in contrast, could help to reduce individual and group defensive routines and thereby more effectively contribute to organizational learning and change (Argyris, 1993; Argyris & Schon, 1978, 1985).

In this regard, we have perhaps the most significant role of allegory as a “way of knowing”—that is, as a way of expressing multiple voices and encouraging multiple interpretations of the change experience. By using allegory, researchers may communicate with participants without directly naming individuals involved in the process, thereby diminishing defensive barriers. Likewise, allegory may be employed to elicit participant responses to the research findings, encouraging those involved to write their own allegory. Such a process could result in an identification of obstacles to change in a way that generates more substantive engagement by members of the organizational community.

**Conclusions**

Allegory, as an extended metaphor, proved to be an excellent framework to better appreciate the symbolic elements of change within the state of Delaware’s organizational community. The interpretive device enabled me to recognize how the diverse meaning systems within the community contributed to human action, and as a result, I became more sensitive to the cultural dimensions in my action research intervention. However, my use of allegory suffered from several key limitations, the most significant of which was in the indirect nature of my presentation of the public text. This yielded inadequate opportunity for the other participants to provide their viewpoints and write their own allegories, and it limited our shared potential to build on the insights as a way of effecting more meaningful, sustainable change.

My limited use of the interpretive device in this particular case should not detract from its potential for future inquiry. Allegory offers participants a way to capture the distinct voices underlying processes of shared change and to recognize the way in which these voices reflect the cultural attributes of their respective organizations. As an approach to action research, allegory also provides a forum to engage participants and to intervene in ways that reduce organizational defensive routines. The literary form may be used as part of a public text in which the participants may be invited to assess themselves and their involvement in the change. In this way, allegory offers promise both as an interpretive device and as a forum for reflexive practice to facilitate the shared-change experience.
References


Joseph W. Grubbs is an assistant professor of public and nonprofit administration at Grand Valley State University in Grand Rapids, Michigan. He completed his Ph.D. in urban affairs and public policy at the University of Delaware. His published work has appeared in the Journal of Organizational Change Management, the JAI Press series, Research in Organizational Change and Development, Public Administration Review, the American Review of Public Administration, the International Journal of Public Administration, the online Journal of Public Administration and Management, and forthcoming in Public Organization Review. He is also a coauthor of the forthcoming book, Public Administration: An Action Orientation, 4th edition (Harcourt Brace).
INDEX

to

ORGANIZATIONAL RESEARCH METHODS

Volume 4

Number 1 (January 2001) pp. 1-88
Number 2 (April 2001) pp. 89-196
Number 3 (July 2001) pp. 197-288
Number 4 (October 2001) pp. 289-396

Authors:


BOIK, CHRISTOPHER, see Bobko, P.

BOIKO, PHILIP, PHILIP L. ROTH, and CHRISTOPHER BOIKO, “Correcting the Effect Size of d for Range Restriction and Unreliability,” 46.

BOIK, ROBERT J., see Aguinis, H.

CABLE, DANIEL M., see Graham, M. E.

CHEN, GILAD, see Cortina, J. M.


DANSEREAU, FRED, see Keller, T.

DUNLAP, WILLIAM P., see Cortina, J. M.

EDEN, DOV, see Chen, G.


FISHER, GWENITH G., see Rogelberg, S. G.


GULLY, STANLEY M., see Chen, G.

HAKEL, MILTON D., see Rogelberg, S. G.

HORVATH, MICHAEL, see Rogelberg, S. G.

HUDY, MICHAEL J., see Wanous, J. P.

LENZ, R. THOMAS, see Anderson, R. D.
MAYNARD, DOUGLAS C., see Rogelberg, S. G.
PIERCE, CHARLES A., see Aguinis, H.
RENSVOLD, ROGER B., see Cheung, G. W.
ROGELBERG, STEVEN G., see Stanton, J. M.
ROTH, PHILIP L., see Bobko, P.
SIMSEK, ZEKI, and JOHN F. VEIGA, “A Primer on Internet Organizational Surveys,” 218.
SMITH, CARLLA S., see Reeve, C. L.
VEIGA, JOHN F., see Simsek, Z.
WILLIAMS, LARRY J., “Editor’s Update,” 193.

Articles:

“Attitudes Toward Surveys: Development of a Measure and Its Relationship to Respondent Behavior,” Rogelberg et al., 3.
“A Community of Voices: Using Allegory as an Interpretive Device in Action Research on Organizational Change,” Grubbs, 376.
“Correcting the Effect Size of $d$ for Range Restriction and Unreliability,” Bobko et al., 46.
“Editor’s Update,” Williams, 193.
“A Generalized Solution for Approximating the Power to Detect Effects of Categorical Moderator Variables Using Multiple Regression,” Aguinis et al., 291.
“Multidimensional Constructs in Organizational Behavior Research: An Integrative Analytical Framework,” Edwards, 144.
“A Primer on Internet Organizational Surveys,” Simsek and Veiga, 218.
“Refining Lodahl and Kejner’s Job Involvement Scale With a Convergent Evidence Approach: Applying Multiple Methods to Multiple Samples,” Reeve and Smith, 91.
“Validation of a New General Self-Efficacy Scale,” Chen et al., 62.