Journal of Applied Psychology
2005, Vol. 90, No. 1, 94–107

Copyright 2005 by the American Psychological Association
0021-9010/05/$12.00 DOI: 10.1037/0021-9010.90.1.94

Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review

Herman Aguinis
University of Colorado at Denver

James C. Beaty
ePredix

Robert J. Boik and Charles A. Pierce
Montana State University

The authors conducted a 30-year review (1969–1998) of the size of moderating effects of categorical variables as assessed using multiple regression. The median observed effect size (f^2) is only .002, but 72% of the moderator tests reviewed had power of .80 or greater to detect a targeted effect conventionally defined as small. Results suggest the need to minimize the influence of artifacts that produce a downward bias in the observed effect size and put into question the use of conventional definitions of moderating effect sizes. As long as an effect has a meaningful impact, the authors advise researchers to conduct a power analysis and plan future research designs on the basis of smaller and more realistic targeted effect sizes.

Using multiple regression to assess the effects of categorical moderator variables (i.e., slope differences across groups) involves a regression equation that examines the relationship between a predictor *X* (e.g., preemployment test scores) and categorical moderator *Z* (e.g., gender) with a criterion *Y* (e.g., a measure of job performance such as supervisory ratings). If the moderator *Z* is binary (i.e., two categories), then the moderated multiple regression (MMR) equation is as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z + \varepsilon, \qquad (1)$$

where β_0 is the intercept, β_1 is the regression coefficient for *X*, β_2 is the regression coefficient for *Z*, β_3 is the regression coefficient for the product term that carries information about the interaction between *X* and *Z*, and ε is a normally distributed random error term (Aguinis, 2004; Cohen, Cohen, West, & Aiken, 2003; Zedeck, 1971). Rejecting the null hypothesis that $\beta_3 = 0$ indicates that *Z* moderates the relationship between *X* and *Y*. Stated differently, the slope of *Y* on *X* differs across values of *Z* (e.g., women and men) and, in this particular example, the preemployment test *X* predicts performance differentially for women and men. Note that although this illustration addresses the binary moderator variable gender, 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/a, and White categories; Aguinis, 2004; West, Aiken, & Krull, 1996).

MMR is the method of choice for testing hypotheses about moderating effects of categorical variables in a variety of research domains such as job performance (e.g., Peters, Fisher, & O'Connor, 1982; Tubbs, 1993), job satisfaction (e.g., Vecchio, 1980), training and development (e.g., Ford & Noe, 1987; Latack, Josephs, Roach, & Levine, 1987), employee turnover (e.g., Bartol & Manhardt, 1979), preemployment testing (e.g., Bartlett, Bobko, Mosier, & Hannan, 1978; Sackett & Wilk, 1994), performance appraisal (e.g., Campion, Pursell, & Brown, 1988), compensation (e.g., Bloom & Milkovich, 1998), organizational citizenship behaviors (e.g., Van Dyne & Ang, 1998), team effectiveness (e.g., Uhl-Bien & Graen, 1998), perceived fairness of organizational practices (e.g., Schaubroeck, May, & Brown, 1994; Tepper, 1994), self-efficacy (e.g., Eden & Zuk, 1995), job stress (e.g., Williams, Suls, Alliger, & Learner, 1991), and career development (e.g., Ibarra, 1995), among others (see Aguinis, 2004, and Stone-Romero & Liakhovitski, 2002, for additional illustrations). In fact, it may be difficult to find a research domain in applied psychology and related fields in which researchers have not tested hypothesized effects of categorical moderator variables.

In spite of the pervasive interest in moderators and the theory-based expectation that moderators should be found, many researchers have concluded that moderating effects do not exist in a number of research domains (e.g., Chaplin, 1997; Schmidt, 1988, 2002; Schmidt & Hunter, 1981; Wigdor & Garner, 1982). However, several simulation studies using Monte Carlo methodology have led to the conclusion that numerous design, measurement, and statistical artifacts bias observed moderating effects downwardly (e.g., Aguinis, 1995; Aguinis & Stone-Romero, 1997;

Herman Aguinis, The Business School, University of Colorado at Denver; James C. Beaty, ePredix, Minneapolis, Minnesota; Robert J. Boik, Department of Mathematical Sciences, Montana State University; Charles A. Pierce, Department of Psychology, Montana State University.

Charles A. Pierce is now at the Department of Management, Fogelman College of Business and Economics, University of Memphis.

Portions of this article were presented at the meetings of the Society for Industrial and Organizational Psychology, New Orleans, LA, April 2000, and Chicago, IL, April 2004.

For additional developments, see the World Wide Web page of Herman Aguinis at http://www.cudenver.edu/~haguinis/

Correspondence concerning this article should be addressed to Herman Aguinis, The Business School, University of Colorado at Denver, Campus Box 165, P.O. Box 173364, Denver, CO 80217–3364. E-mail: herman.aguinis@cudenver.edu

Bobko & Russell, 1994; Russell & Bobko, 1992). Therefore, these artifacts, which are often unavoidable in field settings, may lead researchers to make the sample-based conclusion that there is no moderation when in fact there is a moderating effect in the population.

The failure to detect a moderating effect is an important problem in applied psychology because erroneous decisions have the potential to affect people in very personal ways. For example, erroneous decision making can have particularly detrimental effects on people's careers in such areas as performance management, training and development, and personnel selection. More generally, not detecting moderating effects has detrimental consequences for theory development because researchers may discard incorrectly hypotheses and models including conditional relationships.

Present Study

As noted above, simulation work published mainly in the 1990s has concluded that several design, measurement, and statistical artifacts produce a downward bias in the observed moderating effect size vis-à-vis its population counterpart. In other words, the simulation literature suggests that observed moderating effects in published studies are small. However, at present, there is no comprehensive review of the size of moderating effects in published research. Thus, the first goal of the present study was to answer the following question:

> *Question 1:* What is the size of observed moderating effects of categorical variables in applied psychology and management published research?

Answering this question will allow researchers to understand the magnitude of observed effect sizes across research domains. Also, this information will help researchers understand what may be realistic targeted effect sizes to use in conducting a priori power analyses to help guide the planning of future studies.

A related question is whether the magnitude of observed moderating effect sizes has increased over the past 30 years. The increased sophistication and steady development of theory, together with the routine inclusion of research methodology and statistics courses in graduate programs in psychology and related fields (e.g., Aiken et al., 1990), suggests that the magnitude of moderating effects reported in published research is likely to have increased over time. Stated differently, the improvement in theory regarding the operation of moderator variables, combined with better knowledge regarding efficient research methodology, is likely to have resulted in larger effect sizes in more recently published studies. Thus, a second goal of the present study was to test the following hypothesis:

> *Hypothesis 1:* There will be an increase in the magnitude of observed moderating effects over time.

A third issue relates to the impact of measurement error on the observed moderating effect sizes. Measurement error is one of the determinants of the downward bias in observed effect sizes vis-à-vis population counterparts (Fisicaro & Lautenschlager, 1992). Thus, the third goal of the present study was to answer the following question:

> *Question 2*: What would the size of moderating effects of categorical variables be in applied psychology and management published research if the studies were replicated under conditions in which the predictor *X* and the criterion *Y* have perfect reliability?

Answering this question will allow researchers to learn whether effect sizes computed based on error-free measures are larger than observed moderating effect sizes (which are computed on the basis of fallible measures). We did not expect that construct-level effect sizes (i.e., computed on the basis of error-free measures) would be substantially larger than observed effect sizes because measurement error is only one of several factors that influence effect sizes. Moreover, it is the interactive effects among the various artifacts that produce the largest reduction in effect sizes (Aguinis & Stone-Romero, 1997). In short, although construct-level effect sizes should be larger than observed effect sizes, we did not expect the difference to be substantial.

A fourth issue relates to the ongoing concern about the low power of MMR to detect moderating effects (Aguinis, Boik, & Pierce, 2001). Specifically, there is a need to understand more clearly the relationship among targeted effect sizes, statistical power, and sample sizes reported in published research. Accordingly, a fourth goal of this research was to answer the following question:

> *Question 3*: What is the a priori power of MMR to detect moderating effects of categorical variables in applied psychology and management published research?

Answering this question will allow researchers to understand the relationship between targeted effect sizes and power given the sample sizes reported in published research.

A related fifth goal of the present study was to learn the extent to which applied psychology and management published research has had the ability to detect effect sizes conventionally considered as small, medium, and large (cf. Cohen, 1988). Stated differently,

> *Question 4:* Do MMR tests reported in applied psychology and management published research have sufficient statistical power to detect moderating effects conventionally defined as small, medium, and large?

Answering this question will allow researchers to learn about statistical power values specifically vis-à-vis moderating effect sizes that are conventionally defined as small, medium, and large.

Method

Sample of Studies

We reviewed all articles published from 1969 to 1998 in *Journal of Applied Psychology* (*JAP*), *Personnel Psychology* (*PP*), and *Academy of Management Journal* (*AMJ*). We selected these three journals because they are among the most influential publications devoted to empirical research in applied psychology and management (Starbuck & Mezias, 1996). In addition, these journals have a reputation of enforcing the highest methodological standards. Thus, the resulting effect sizes should be liberal. Stated differently, we expected that given the methodological rigor and emphasis on theory of *AMJ*, *JAP*, and *PP*, the effect sizes of studies published in these outlets should be as large or larger than those of studies published in other applied psychology and management journals. The criteria for including studies in the review were as follows: (a) at least one MMR analysis was included as part of the study, (b) the MMR analysis included a continuous criterion, (c) the MMR analysis included a continuous predictor, and (d) the MMR analysis included a categorical moderator.

Study Identification and Accuracy Checks

James C. Beaty located all relevant studies published between 1969–1998 by performing a manual search of each issue of *AMJ*, *JAP*, and *PP* using the criteria outlined above. This search resulted in a total of 106 articles. The list of these articles can be obtained by contacting Herman Aguinis. Virtually every article reported more than one MMR analysis. The total number of reported MMR analyses was 636.

Herman Aguinis conducted a random check of 10 journal issues to assess the completeness of the review and statistics extracted from the articles. The check resulted in 100% agreement with results obtained by James C. Beaty.

Computation of Effect Size and Statistical Power

General considerations. As expected, the majority of the published articles did not include all the information needed to compute effect sizes. The vast majority of articles reported sample size across moderator-based subgroups. However, fewer reported predictor–criterion correlations, and even fewer reported information regarding within-group variances. We conducted a systematic effort to contact each of the authors to obtain additional statistics not included in the articles. Contact information for authors was obtained from the following sources: (a) information provided in the article, (b) Society for Industrial and Organizational Psychology 1999 membership directory, and (c) Academy of Management membership directory (available online at http://www.aomonline.org/). As noted above, our review identified a total of 106 articles, including 636 MMR analyses. Of this total number of MMR analyses, after contacting authors directly, we had information regarding sample size across moderator-based subgroups for 507 (79.72%) analyses, information regarding sample size and predictor–criterion correlations across moderator-based subgroups for 261 (41.04%) analyses, information regarding the variance of the predictor *X* within moderator-based subgroups for 151 (23.74%) analyses, and information regarding the variance of the criterion *Y* within moderator-based subgroups for 173 (27.20%) analyses.

Although there are interactive effects, sample size and predictor–criterion relationships across moderator-based subgroups have been identified as the two most influential factors on the observed effect size (cf. Aguinis et al., 2001; Aguinis & Stone-Romero, 1997). Also, correlations based on observed scores reflect the impact of other factors known to affect parameter estimates such as measurement error and expected ratio of sample variance to population variance (i.e., truncation). Thus, including the 261 MMR analyses for which information regarding sample size and correlations is available results in an empirical estimate of the distribution of effect sizes, but the precise manner in which these values have been affected by factors known to influence effect sizes (e.g., measurement error, truncation) remains unknown.

An additional consideration is that effect sizes may be affected by whether the moderator was the focus of the study. For example, a researcher may test and report results of an MMR analysis even though it is not a focus of the research (e.g., to test for equal slopes before conducting a covariance analysis). And, hypothesized moderating effects are likely to be larger than nonhypothesized moderating effects. Thus, James C. Beaty reviewed the description of each of the 261 MMR analyses and coded them as to whether each moderator test was specifically hypothesized. There were fewer than 10 tests for which there was not certainty as to whether the test was the focus of the research, so Herman Aguinis also read the description for each of these analyses independently. There was complete agreement between Herman Aguinis and James C. Beaty regarding these MMR tests. The result of this content analysis yielded 257 (98.5%) tests that were specifically hypothesized. Given that virtually all tests were hypothesized and it would not be meaningful to compare whether the set of 257 hypothesized effects is larger than the set of 4 nonhypothesized effects, we decided to conduct the analyses based on the entire set of 261 tests.

Computation of effect size. In the case of MMR, an effect size metric that can be used across diverse studies and measurement scales is f^2. This measure of effect size describes the strength of the moderating effect. Specifically, f^2 is the ratio of systematic variance accounted for by the moderator relative to unexplained variance in the criterion (Aiken & West, 1991). Aiken and West (1991, p. 157) described how to compute f^2 for the case of a continuous moderator variable under the assumption of homogeneity of error variance. Their equation is not appropriate, however, if the homogeneity of error variance assumption is violated. Accordingly, we developed a modified f^2 that is appropriate for situations with categorical moderator variables when there is heterogeneity of error variance. The derivation of this modified f^2 is an additional unique contribution of the present study and is included in Appendix A. Although the equations are complex, a computer program available at http://carbon.cudenver.edu/~haguinis/mmr/ performs all needed computations online.

To compute construct-level f^2, we recomputed f^2 for a hypothetical replication of the study conducted under conditions in which *X* and *Y* have perfect reliability. Accordingly, the value of $\rho_{\text{observable}}$ (i.e., correlation between the observable scores for predictor *X* and the criterion *Y* in each moderator-based subpopulation) is larger in the error-free replication than it is in the original study, and this is why f^2 is expected to be larger. In other words, we asked the following question: How would the properties of the study (i.e., effect size f^2) change if the study was replicated under conditions in which *X* and *Y* have perfect reliability?

Note that Bobko and Rieck (1980) and Raju and Brand (2003) examined standard errors of sample correlations corrected for measurement error as well as properties of statistical tests based on the corrected correlations. Those studies revealed that tests of population correlations that are based on corrected sample correlations are, in general, no more powerful than are tests based on uncorrected sample correlations. In the present study we ask a slightly different question. Our question is, How large of an impact does measurement error have on effect sizes? That is, if two studies are equivalent in all respects except for measurement error, then what is the expected difference in their effect sizes? Our question concerns the impact of measurement error, not the impact of correcting for measurement error.

Finally, we did not have reliability information for all the MMR tests. The mean reliability for the 46 tests (i.e., 18%) for which this information was available for *X* was .80, and the mean reliability for the 50 tests (i.e., 19%) for which this information was available for *Y* was .81. Thus, we computed construct-level effect sizes using the reported reliability when available and used a value of .80 for the remaining effect sizes. Appendix B includes a technical description of the computation of construct-level effect sizes.

Computation of statistical power. Aguinis and colleagues (Aguinis & Pierce, 1998b; Aguinis, Pierce, & Stone-Romero, 1994) developed computer programs to estimate the power of MMR. These programs, written using empirically based algorithms, allow researchers to estimate the power of MMR tests for given values for factors known to affect power (e.g., moderating effect magnitude, sample size across moderator-based groups). Despite the fact that these programs are available and aid researchers in the quest for moderating effects, Aguinis et al. (2001) noted that they suffer from four 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; Aguinis et al., 1994, is based on Stone-Romero, Alliger, & Aguinis, 1994). First, these programs do not include all the factors known to affect the power of MMR. Second, these programs are based on Monte Carlo studies that included only a limited range of values for factors affecting the power of MMR. Third, these programs assume that restriction on the continuous predictor *X* takes on only the simplest form of variance reduction (i.e., truncation). Fourth, users of the programs can only compute power in situations in which the categorical moderator has two levels.

To overcome the aforementioned limitations of empirically derived algorithms to compute power of MMR to detect categorical moderator variables, Aguinis et al. (2001) developed a theory-based approximation. Therefore, in the present study we used the Aguinis et al. (2001) theorem to compute power corresponding to a range of f^2 values. This approximation is included in Appendix C (see Aguinis et al., 2001, for a detailed discussion and development of this theorem).

The effects of violating MMR's homogeneity of error variance assumption on statistical power are complex and include a possible inflation of Type II error rates (i.e., a decrease in power) as well as, in some cases, inflation of Type I error rates (Aguinis, 2004; Aguinis & Pierce, 1998a; Alexander & DeShon, 1994). In computing power, we used *Y* variance information for the 173 out of 261 (66.28%) MMR analyses for which this information was available. For those analyses for which information regarding *Y* variances in moderator-based subgroups was not available, we assumed homogeneity of error variance.

It should be noted that several authors have argued against the usefulness of assessing power based on observed effect sizes (i.e., post hoc power analysis; e.g., Gerard, Smith, & Weerakkody, 1998; Goodman & Berlin, 1994; Hoenig & Heisey, 2001). For example, Hoenig and Heisey (2001) demonstrated that power based on observed effect sizes is a direct function of the obtained *p* value for the test in question and that this power value cannot provide more information than the reported *p* value. Consequently, once data are collected, a power analysis provides no additional information beyond that contained in the confidence interval (CI) around the parameter estimate of interest. To illustrate this general principle, consider the following scenario. A researcher conducts a study and finds that $f^2 = .0001$. That is, the ratio of explained variance by the moderator to unexplained variance in the criterion is only one hundredth of one percent. Conducting a post hoc power analysis based on this observed effect size would lead to the conclusion that power is abysmal, unless sample size is very large. But, by doing so, the researcher would be asking the question, What would the power be to detect an effect size of .0001? Most likely, researchers would not be interested in knowing the power for such a small targeted effect size. An alternate explanation for the low power in this situation is that a nonsignificant interaction was found because the population moderating effect is small. Thus, the logical flaw in conducting a post hoc power analysis is to posit that whatever effect size was observed—no matter how small—is one a researcher would wish to find statistically significant. As noted by an anonymous reviewer, one can certainly perform the exercise of determining the power to detect observed effect sizes, but finding low power in such a setting may not be a cause for alarm.

Given the above considerations, we did not compute power based on observed effect sizes. Instead, we computed power for each published MMR test included in our review for targeted values for f^2 ranging from a low of .001 to a high of .35 (i.e., a ratio of variance explained by the moderator to unexplained variance ranging from 0.01% to 35.00%). Also, in computing f^2 for each published MMR test, we used the sample sizes reported in each article. The chosen range of values for f^2 is likely to include critical effect sizes (i.e., effect sizes considered of practical and/or scientific importance) for most applied psychology and management research areas. Note that the criticality of a specific magnitude for f^2 must be evaluated against the area of research considered, the anticipated research outcomes, subsequent impact, and other factors. And, an effect size of .001 may not be sufficiently large to warrant detection in most research domains. However, we chose to include a broad range of targeted effect sizes for the sake of completeness.

In summary, we computed effect sizes for the 261 MMR analyses for which sample size and predictor–criterion correlations across moderator-based subgroups were available. We used *X* and *Y* variance information when available and assumed homogeneity of error variance for the MMR analyses for which this information was not available. In computing the construct-level effect sizes (i.e., based on error-free measures), we used reliability information when available and a value of .80 when this information was not available. In computing power, we used targeted f^2 values ranging from .001 to .35 to include a broad range of what can be considered critical effect sizes in applied psychology and management research and the sample sizes reported in each published study.

Results

Frequency of MMR Use Over 30-Year Review Period

Figure 1 shows the number of MMR analyses published in *AMJ*, *JAP*, and *PP* over the 30-year period covered in our review (i.e., 1969–1998). As can be seen in Figure 1, the use of MMR to assess categorical moderators has remained fairly stable at approximately 20–40 analyses per year since the mid-1980s.

Question 1: Size of Observed Moderating Effects

Overall effect sizes. Table 1 shows that the overall mean observed effect size (i.e., f^2) for the 261 analyses is .009, with a 95% CI ranging from .0089 to .0091. However, because the distribution of effect sizes is positively skewed (skewness = 6.52, $z = 21.73$, $p < .01$), the median effect size of .002 is a better descriptor of central tendency. The 25th percentile is .0004 and the 75th percentile is .0053.

Comparison across journals. Next, we computed observed effect sizes for *AMJ* ($k = 6$), *JAP* ($k = 236$), and *PP* ($k = 19$). For *AMJ*, the mean and median effect sizes are .040 and .025 ($SD = .047$); for *JAP*, these mean and median values are .007 and .002 ($SD = .024$); and for *PP*, these mean and median effect sizes are .017 and .006 ($SD = .025$). Because of the high degree of skewness, we normalized the distribution of effect sizes by implementing the Box–Cox family of power transformations (Box & Cox, 1964). Specifically, observed effect sizes were transformed by using the .15 root, that is $(f^2)^{.15}$, which resulted in an approximately normal distribution (i.e., skewness = .25, kurtosis = .63). We then conducted an analysis of variance (ANOVA) on the transformed effect sizes to examine possible differences across the three journals. Results showed a statistically significant difference, $F(2, 258) = 8.71$, $p < .001$, $\eta^2 = .06$. Tukey's honestly significant difference (HSD) tests showed that effect sizes reported in *AMJ* are larger than those reported in *JAP* ($p = .002$) and those reported in *PP* were also larger than those reported in *JAP* ($p = .028$), but there was not a statistically significant difference between the mean effect size reported in *AMJ* versus *PP*.

Comparison across moderator type. Of the total 261 MMR analyses for which we computed effect sizes, 63 addressed the moderating effect of gender, 45 addressed the moderating effect of ethnicity, and 153 addressed other moderators. For tests including gender, Table 1 shows that the mean and median effect sizes are .005 and .002 ($SD = .011$). For tests regarding ethnicity, the mean and median effect sizes are .002 and .001 ($SD = .002$). For other moderators, the mean and median effect sizes are .013 and .002 ($SD = .031$). An ANOVA based on the transformed effect size values showed a statistically significant difference across the three groups, $F(2, 258) = 4.97$, $p = .008$, $\eta^2 = .04$. Tukey's HSD tests showed that effect sizes for tests of ethnicity as a moderator are

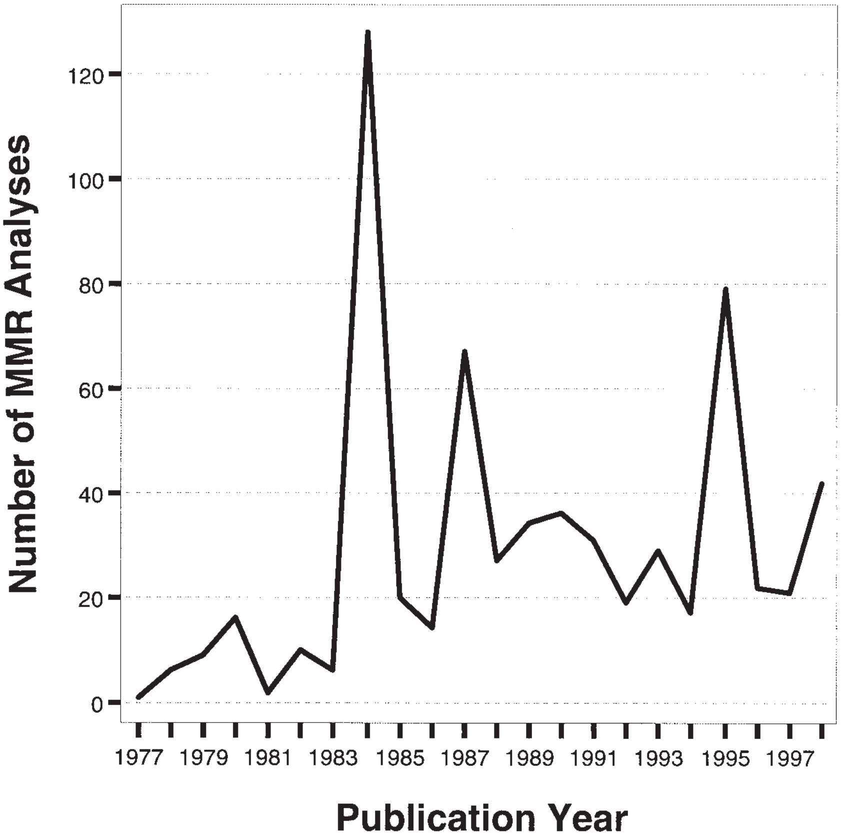

Figure 1. Number of moderated multiple regression (MMR) analyses of categorical moderator variables reported in *Academy of Management Journal*, *Journal of Applied Psychology*, and *Personnel Psychology* between January 1969 and December 1998. Frequencies for 1969–1976 = 0.

smaller than effect sizes reported for the “other” category (p = .005), but there were no other differences.

Comparison across research domains. Next, we investigated whether effect sizes differ across research domains. First, we compared personnel selection versus other areas. Second, we compared the work attitudes literature versus other areas.

Personnel selection versus other areas. Of the total 261 MMR analyses for which we computed effect sizes, 20 specifically addressed a personnel selection issue, whereas 241 addressed other topics. For tests in the personnel selection domain, the mean and median effect sizes are .010 and .001 (SD = .023). For tests in other research domains, the mean and median effect sizes are .009 and .002 (SD = .025). An independent-samples t test using the transformed effect sizes as the dependent variable showed no statistically significant differences between personnel selection and other research domains, $t(259) = -0.23$, $p = .82$.

Work attitudes versus other areas. Of the total 261 analyses, 96 tested categorical moderators with work attitudes (e.g., job satisfaction, organizational commitment) as a criterion. For moderator tests in the work attitudes domain, the mean and median effect sizes are .005 and .002 (SD = .015). For tests not using work attitudes as criteria, the mean and median effect sizes are .009 and .002 (SD = .025). Results of an independent-samples t test showed no statistically significant differences for the normalized effect sizes comparing work attitudes with other research areas, $t(259) = -0.95$, $p = .34$.

Hypothesis 1: Observed Effect Sizes Over Time

To test the hypothesis that moderating effects have increased in magnitude over time, we computed Pearson’s correlation coefficient between year of publication and effect size. Results showed a statistically significant relationship such that $r(261) = .15$, $p < .05$. That is, more recently published studies generally report larger effect sizes than older studies.

Question 2: Size of Construct-Level Moderating Effects

Table 1 shows summary statistics for the construct-level (i.e., based on error-free measures) effect sizes. Comparing results of observed versus construct-level effect sizes indicates that the use of error-free measures increased the overall effect size from a median of .002 to .003 and from a mean of .009 to .017. Thus, if X and Y are measured error-free, the median effect size increased only by .001. Although this represents an increase of 50%, it is small in absolute terms. This was expected, given that measurement error is only one of the several design, measurement, and statistical artifacts that influence effect sizes and that these artifacts

Table 1
Summary Statistics for Observed Effect Sizes, Construct-Level Effect Sizes, and Sample Sizes Across Journals, Type of Moderator, and Research Domains (1969–1998)

	Observed effect size (f^2)			Construct-level effect size (f^2)			*n*	
Comparison	*Mdn*	*M* (*SD*)	95% CI	*Mdn*	*M* (*SD*)	95% CI	*Mdn*	*M* (*SD*)
Journal								
AMJ (k = 6)	.025	.040 (.047)	.0397–.0403	.044	.067 (.082)	.0665–.0675	45	293 (652)
JAP (k = 236)	.002	.007 (.024)	.0069–.0071	.002	.014 (.049)	.0138–.0142	158	402 (559)
PP (k = 19)	.006	.017 (.025)	.0166–.0174	.009	.040 (.068)	.0389–.0411	101	122 (92)
Moderator type								
Gender (k = 63)	.002	.005 (.011)	.0049–.0051	.003	.011 (.041)	.0107–.0113	245	230 (208)
Ethnicity (k = 45)	.001	.002 (.002)	.0020–.0020	.002	.003 (.004)	.0030–.0030	245	1006 (968)
Other (k = 153)	.002	.013 (.031)	.0128–.0132	.003	.024 (.062)	.0235–.0245	90	252 (270)
Research domain (*I*)								
Personnel selection (k = 20)	.001	.010 (.023)	.0097–.0103	.002	.029 (.067)	.0281–.0299	88	153 (128)
Other (k = 241)	.002	.009 (.025)	.0089–.0091	.003	.016 (.050)	.0158–.0162	158	396 (563)
Research domain (*II*)								
Work attitudes (k = 96)	.002	.005 (.015)	.0049–.0051	.003	.010 (.039)	.0098–.0102	351	311 (334)
Other (k = 165)	.002	.011 (.029)	.0109–.0111	.003	.021 (.058)	.0207–.0213	158	416 (632)
Overall (k = 261)	.002	.009 (.025)	.0089–.0091	.003	.017 (.052)	.0167–.0173	158	378 (545)

Note. Construct-level f^2 is computed based on error-free measures for X and Y (see Appendix B for computational details). *AMJ* = *Academy of Management Journal*; *JAP* = *Journal of Applied Psychology*; *PP* = *Personnel Psychology*; k = number of moderator tests; f^2 = ratio of systematic variance accounted for by the moderating effect relative to unexplained variance in the criterion (see Appendix A for computational details); n = sample size for moderator-based subgroups; CI = confidence interval.

have interactive effects (Aguinis & Stone-Romero, 1997). An examination of the specific comparisons shown in Table 1 reinforces the conclusion that, overall, the impact of measurement error on the absolute magnitude of effect sizes is not substantial.

Question 3: Statistical Power

Table 1 shows that the overall median moderator-based subgroup sample size is 158. This median sample size is larger than the median sample size of 113 reported by Salgado (1998) for criterion-related validity studies in the personnel selection domain published between 1983 and 1994 in *JAP*, *PP*, and *Journal of Occupational and Organizational Psychology*. But, sample sizes in the present review are not inordinately large as compared to those reported in Salgado's review.

Table 2 includes summary statistics for power values associated with a broad range of targeted effect sizes given the reported sample sizes in the studies included in our review (Appendix C includes computational details regarding power). Results show that effect sizes do not need to be too large to be detected. Specifically, results in Table 2 indicate that power in published research is sufficient (i.e., M = .84, *Mdn* = .90) to detect an effect size of .02. Moreover, 72% of the tests reviewed had power of .80 or greater to detect an effect of .02.

Table 2 also includes information on what percentage of tests achieved power of at least .80 for each targeted f^2 value. As shown in this table, only 21.80% of tests included in our review had power of at least .80 to detect f^2 =.01, but 85.80% of tests had power of .80 or greater to detect f^2 = .03, and each test had power of .80 or greater to detect f^2 = .35.

Figure 2A shows a graph depicting the relationship between effect size and mean power based on the data presented in Table 2 but limited to the .001–.10 f^2 range only (it would be redundant to include values greater than f^2 = .10, because the resulting mean

Table 2
Statistical Power to Detect Targeted Effect Sizes in Published Applied Psychology and Management Research

	Power		
Targeted effect size (f^2)	*Mdn*	*M* (*SD*)	% above .80
.001	.11	.14 (.11)	0.40
.002	.17	.22 (.16)	0.80
.003	.24	.29 (.20)	6.10
.004	.30	.35 (.22)	8.80
.005	.36	.41 (.23)	8.80
.006	.42	.46 (.24)	11.10
.007	.48	.50 (.25)	13.80
.008	.53	.54 (.25)	16.50
.009	.58	.58 (.25)	17.60
.010	.62	.61 (.25)	21.80
.020	.90	.84 (.21)	72.00
.030	.98	.90 (.21)	85.80
.040	1.00	.93 (.19)	88.90
.050	1.00	.94 (.17)	89.30
.060	1.00	.95 (.15)	90.80
.070	1.00	.96 (.14)	93.50
.080	1.00	.96 (.13)	93.50
.090	1.00	.97 (.12)	93.50
.100	1.00	.97 (.11)	93.90
.120	1.00	.98 (.10)	95.40
.140	1.00	.98 (.09)	96.60
.160	1.00	.99 (.07)	97.70
.180	1.00	.99 (.06)	98.50
.200	1.00	.99 (.06)	98.50
.250	1.00	.99 (.04)	98.50
.300	1.00	1.00 (.03)	98.50
.350	1.00	1.00 (.02)	100.00

Note. f^2 = ratio of systematic variance accounted for by the moderating effect relative to unexplained variance in the criterion (see Appendix A for computational details regarding f^2 and Appendix C for computational details regarding power); % above .80 = percentage of moderated multiple regression analyses (out of a total of 261) with a power value greater than .80.

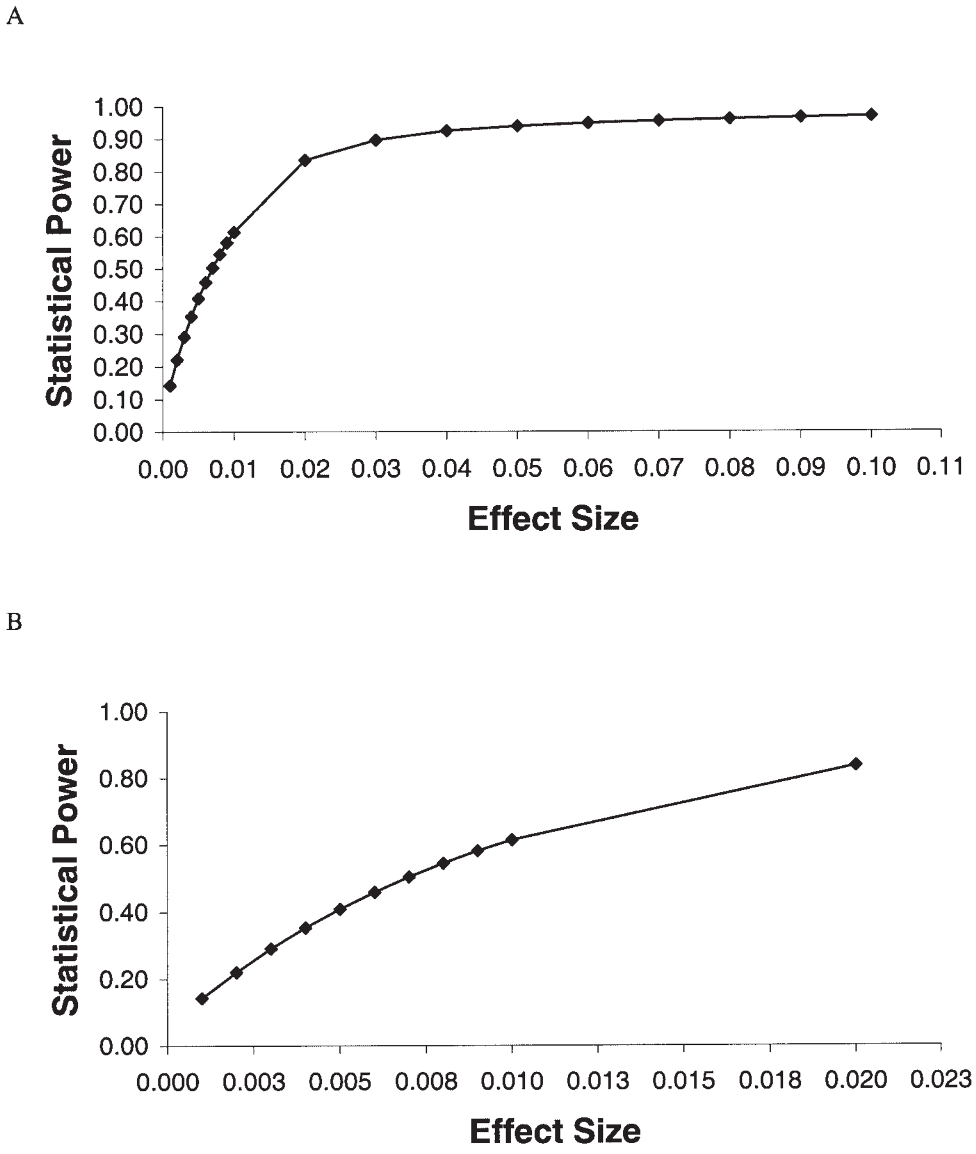

Figure 2. Graphical representation of the relationship between targeted effect size (f^2) and power for the .001 to .10 f^2 range (A) and the .001 to .02 f^2 range (B).

power is greater than .97). This graph shows that the moderator test achieves a mean power of .80 for a targeted f^2 value of just under .02. And, mean power reaches .90 for a targeted f^2 value of .03. Figure 2B shows a more detailed graph of the relationship between effect size and mean power for the lower range of the f^2 continuum (i.e., .001 to .02). This graph illustrates that, at this low end of the f^2 range, power increases very rapidly with small increments in the targeted effect size value. For example, mean power increases from .17 to .62 as targeted f^2 values increase from .001 to .01 (i.e., a change of only .09% in the ratio of variance explained by the moderator to unexplained variance in the criterion).

Appendix D includes an analytic description of how to determine f^2 corresponding to a specific power and sample size. The procedure described in Appendix D involves first determining the noncentrality parameter λ. Then, f^2 is computed as follows (this approximation becomes exact only under restrictive assumptions):

$$f^2 \approx \frac{2\lambda}{N - 2k}, \tag{2}$$

where k is the number of moderator-based subgroups and N is the total sample size.

We implemented the calculations shown in Appendix D for the median moderator-based subgroup sample size of 158. Given a test including two moderator-based subgroups, results show that power would reach .80 for a targeted $f^2 = .0253$. This result is consistent with the finding discussed above that f^2 must be approximately .02 to achieve power of .80 (also, see Figure 2A).

Question 4: Power to Detect Conventionally Defined Small, Medium, and Large Effect Sizes

A number of authors (e.g., Aiken & West, 1991) have echoed Cohen's (1988) conventional definitions of small, medium, and large effect sizes. For the case of f^2, Cohen (1988) suggested that effect sizes around .02, .15, and .35 be labeled *small*, *medium*, and *large*, respectively. It should be noted that Cohen (1988; see also Cohen et al., 2003) offered the caveat that even effect sizes labeled *small* can have substantial practical and theoretical importance. Nevertheless, Cohen's conventional definitions are used pervasively, particularly in literature reviews of statistical power (e.g., Brock, 2003; Mazen, Graf, Kellogg, & Hemmasi, 1987; Mazen, Hemmasi, & Lewis, 1987).

As noted above, Table 2 shows that the mean power of the MMR test to detect what is conventionally defined as a small effect (i.e., $f^2 = .02$) is .84. The mean power to detect a medium effect (i.e., $f^2 = .15$) is approximately .98, and the power to detect a large effect (i.e., $f^2 = .35$) is 1.0.

Discussion

Given that simulation studies have demonstrated that numerous design, measurement, and statistical artifacts produce a downward bias in the magnitude of observed moderating effects, we expected that observed effect sizes would be quite small. We had this expectation despite the fact that literally hundreds of researchers in dozens of disparate domains in applied psychology, management, and associated fields have hypothesized such effects. Our results show that the median effect size is .002. And, the effect size "bandwidth" is uniformly narrow and around .002 for the areas of personnel selection and work attitudes and for tests including the moderating effect of gender and ethnicity. Also, computations of the moderating effect size based on error-free measures increased the size of the median moderating effect by only .001.

In spite of the result that observed moderating effects are small, the present review suggests a number of more encouraging results. First, none of the 95% CIs around the mean effect size for the various comparisons shown in Table 1 include the value of zero. Second, observed moderating effects have increased in magnitude over time, albeit this positive trend is not that strong because the correlation between year of publication and observed effect size is only .15. Third, our results suggest that, given the sample sizes reported in the studies reviewed, statistical power has, in general, been sufficient (i.e., .80 or greater) to detect effects of a magnitude of .02 or greater. And, 72% of the tests reviewed had a statistical power of .80 or greater to detect effects of at least $f^2 = .02$. Fourth, our results indicate that if an MMR test includes a moderator with two categories (e.g., gender) and subgroup sample sizes of 158, power would be .80 for a targeted f^2 of approximately .02. Finally, regarding Cohen's (1988) definitions of small ($f^2 = .02$), medium ($f^2 = .15$), and large ($f^2 = .35$) effect sizes, our results indicate that approximately 72% of the tests reviewed had sufficient power (i.e., .80 or greater) to detect a small effect, approximately 85% of tests had sufficient power to detect a medium effect, and 100% of tests had sufficient power to detect a large effect. In sum, although observed effect sizes are substantially smaller than expected, statistical power is sufficient to detect what is conventionally defined as a small targeted effect size.

Implications for Theory

As shown in Figure 1, the use of MMR to assess the moderating effects of gender, ethnicity, and other categorical variables is pervasive. In fact, there are few areas, if any, in applied psychology and management that are not researched using MMR to test hypotheses regarding categorical moderators of a wide diversity of variables such as gender, ethnicity, goal difficulty, organizational unit, type of feedback, task control, training method, employment status, type of compensation, leadership style, nationality, ownership control, pay plan, and so forth. Given the small median effect size of .002 revealed by our review, could it be that theories in all of these research domains are incorrect in positing the operation of moderators? Could it really be the case that literally hundreds of researchers in dozens of disparate domains in applied psychology and associated fields are wrong and that population moderating effects are of such small magnitude? We cannot discard this possibility. However, it seems more likely that the pervasive and often unavoidable design, measurement, and statistical artifacts decrease the observed effect sizes substantially vis-à-vis their population counterparts. In fact, this seems to be a more plausible explanation given the Monte Carlo evidence demonstrating the impact of such artifacts including their interactive effects. Accordingly, given the pervasive use of MMR, it is likely that numerous hypotheses regarding moderating effects have been discarded incorrectly over the past 30 years. Thus, an important implication for theory development is that past failures to find support for hypothesized moderators may have been due to observed effect sizes being smaller than their population counterparts. Consequently, we suggest that past null findings be closely scrutinized to assess whether they may have been due to the impact of artifacts as opposed to the absence of a moderating effect in the population.

Implications for Research and Organizational Practices

The present study has several implications for the conduct of research aimed at assessing moderating effects. First, we urge researchers to be more sensitive to the methodological and statistical artifacts known to produce a downward bias in the observed effect size (see Aguinis, 2004, Chapter 5, for a detailed description of each of these artifacts). More attention to research design issues is likely to lead to a considerable payoff in terms of increasing the observed effect size, and consequently, the likelihood that population effects will be detected. This is particularly true given that, for typical observed sample sizes, power reaches .80 for a targeted effect size of approximately only .02. But, researchers may not be able to observe a targeted effect size of this magnitude unless careful attention is given to design and measurement issues.

Second, we suggest that researchers become more aware of the factors that affect the power of MMR and implement recently developed computer programs to calculate the power of MMR in planning a study's design. Such programs are in the public domain and descriptions, as well as instructions on how to obtain them, can be found in Aguinis (2004); Aguinis and Pierce (1998b); Aguinis, Petersen, and Pierce (1999); and Aguinis et al. (1994, 2001).

Third, the present results provide information to help evaluate the appropriateness of Cohen's (1988) traditional definitions of small, medium, and large effect size in conducting power analyses regarding moderating effects of categorical variables. First, we emphasize that the choice for a targeted effect size in a power analysis should not be based on broad-based conventions but rather on the specific research situation in hand. Cohen (1977) himself made this recommendation almost 30 years ago in his power analysis book when he stated that effect size distinctions are relative "to the area of behavioral science or even more particularly to the specific content and research methods being employed" (p. 25). And, similarly, a recent editorial in *AMJ* (Eden, 2002) stated that "the importance of any particular effect size depends upon the nature of the outcome studied" (p. 845). In spite of this, many researchers have used Cohen's definitions of small, medium, and large effect sizes in conducting power analyses (e.g., Mone, Mueller, & Mauland, 1996; Sedlmeier & Gigerenzer, 1989) and, moreover, have not acknowledged explicitly that these values are actually based on observed effect sizes for specific, and often limited, literature domains. Specifically, the values derived by Cohen (1962) are based on observed effect sizes computed from articles published in just one volume of *Journal of Abnormal and Social Psychology*. More precisely, in his Method section, Cohen noted that "the level of average population proportion at which the power of the test was computed was the average of the sample proportions found" and "the sample values were used to approximate the level of population correlation of the test" (p. 147). And, because Cohen's now conventional definitions of small, medium, and large effect sizes are based on observed values, they have been revised over time as a consequence of subsequent literature reviews of effect sizes in various domains. For example, for correlation coefficients, Cohen defined .20 as small, .40 as medium, and .60 as large in his 1962 *Journal of Abnormal and Social Psychology* review. However, he changed these definitions to .10, .30, and .50 in his 1988 power analysis book. The present review shows that even if a researcher hypothesizes what can be conventionally considered a "small" moderating effect size (i.e., $f^2 = .02$), and moreover, plans the research design accordingly so that power will be .80 to detect an effect of .02 or larger, the moderating effect may not be found given that the median observed effect found in our review is .002. Note, however, that the finding that the median effect size is only .002 does not necessarily suggest that this is the targeted value that should be used in computing power. True effects are not necessarily important effects and the targeted value should be chosen on the basis of the anticipated impact of the expected effect size for theory and/or practice. Nevertheless, our results show that as long as a small effect has a meaningful impact for science or practice within a specific context, the implication is that researchers should conduct a power analysis and plan future research designs based on smaller (and more realistic) targeted effect sizes as opposed to Cohen's (1962) conventional definitions, which are largely based on a review of articles published in just one volume of one journal that did not include research from the applied psychology and management fields.

The present study also has implications for organizational practices. For instance, assume that a particular intervention involving participative decision making has a positive effect on a specific group's performance (e.g., Generation Xers) and a negative effect on another group's performance (e.g., Baby Boomers). Not detecting this moderating effect of group membership may lead management to decide incorrectly that the intervention should be implemented for all employees. Numerous additional examples can be used to show that the inability to recognize conditional relationships, including those with group membership as a moderator, may lead to decision making that results in detrimental consequences for both individuals and organizations. For example, not detecting a moderating effect of, for instance, gender or ethnicity may lead to incorrect selection decisions. Moreover, it may lead to making hiring decisions that penalize certain applicants on the basis of group membership. Eventually, not recognizing the presence of a moderator variable in the selection context may lead to group-based differences in performance scores that can lead to costly lawsuits.

Limitations

We close by discussing limitations of the present study. First, we reviewed articles published in only three journals. Arguably, we could have reviewed other publications in applied psychology and management. However, our reasoning was that if effect sizes are small in three of the most methodologically rigorous journals, they will be at least as small in other publications. Thus, we speculate that the present results are actually an overestimate of the effect sizes that would be found combining *AMJ*, *JAP*, and *PP* with other journals.

Second, a number of published articles did not include sufficient information to compute effect sizes. This is a problem that is mentioned frequently by researchers who conduct other types of quantitative literature reviews such as meta-analysis. Therefore, we contacted individual authors to obtain additional information. Much to our surprise, the vast majority of authors had not kept their data or did not have access to them. Given this situation, we could only compute effect sizes for 261 of the 636 MMR analyses (i.e., 41.04%). Although we computed effect sizes for fewer than half of all the analyses reported over the 1969–1998 period, we do not have any reasons to believe that the sample of 261 analyses is not representative of the total population of 636.

Third, one may ask whether researchers should even bother conducting a power analysis and attempting to detect a moderating effect expected to be, for example, no greater than $f^2 = .01$. There is no general answer to this question because, as noted above, the specific research question within a specific research domain dictates whether a specific effect size is practically or scientifically important. For example, the fact that even small effect sizes can be of high practical or scientific importance was illustrated by Martell, Lane, and Emrich (1996). Specifically, an effect size of 1% regarding male–female differences in performance appraisal scores led to only 35% of the highest level positions being filled by women. On the basis of these results, Martell et al. concluded that "relatively small sex bias effects in performance ratings led to substantially lower promotion rates for women, resulting in proportionately fewer women than men at the top levels of the organization" (p. 158). Several additional illustrations of the impact of what may be conventionally considered "small" effects on science and practice are provided by Breaugh (2003), Fichman (1999), and Olejnik and Algina (2000).

Fourth, our methodology for computing f^2 included information relating to the continuous predictor and categorical moderator only. Some studies included additional predictors (e.g., control variables) in the MMR models. For example, Van Dyne and Ang (1998) entered five categorical control variables into the equation before the continuous predictor and categorical moderator and two had statistically significant regression coefficients. However, it was not possible to compute f^2 after all predictors are included in the regression model. This was the case because we would need, at a minimum, the correlation matrix including all predictors and the criterion and we would need this correlation matrix for each moderator-based subgroup. This information is not available in the published articles and, given our experience in trying to obtain additional information directly from authors, it would be virtually impossible to gather the necessary data to perform the calculations. Specifically, values for f^2 may change if additional variables were included in the model because the addition of predictors is likely to decrease the amount of unexplained variance in the criterion. Because f^2 is expressed as a ratio of explained to unexplained variance, the inclusion of additional predictors may increase the value for f^2 (assuming a constant relationship between the moderator and the criterion). In spite of this, the substantive conclusion that moderating effect sizes in published research are small is not likely to change even if we had information regarding the contribution of additional predictors. For example, assume the extreme situation in which the inclusion of additional predictors decreased unexplained variance fivefold. In this rather extreme scenario, and assuming a constant relationship between the moderator and the criterion, the observed median effect size would be .01, which is still half the size of what is currently considered a "small" effect (Cohen, 1988).

Concluding Comments

In closing, our review documents that, overall, observed moderating effects are smaller than what is conventionally defined as a small effect (cf. Cohen, 1988). As long as such small effects have a meaningful impact for science or practice within a specific context, we advise researchers to conduct a power analysis and plan future research designs on the basis of smaller, more realistic, and yet meaningful targeted effect sizes. The advantage of power analyses that use more realistic (but nevertheless important) targeted effect sizes is that the resulting research designs should be more conducive to detecting hypothesized effects. In addition, researchers are advised to follow recommendations on how to minimize the impact of design, measurement, and statistical artifacts that have a downward bias on the effect size. Following these strategies is likely to allow researchers to identify meaningful moderated relationships more effectively.

References

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management, 21,* 1141–1158.

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford Press.

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods, 4,* 291–323.

Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods, 2,* 315–339.

Aguinis, H., & Pierce, C. A. (1998a). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods, 1,* 296–314.

Aguinis, H., & Pierce, C. A. (1998b). Statistical power computations for detecting dichotomous moderator variables with moderated multiple regression. *Educational and Psychological Measurement, 58,* 668–676.

Aguinis, H., Pierce, C. A., & Stone-Romero, E. F. (1994). Estimating the power to detect dichotomous moderators with moderated multiple regression. *Educational and Psychological Measurement, 54,* 690–692.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82,* 192–206.

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

Aiken, L. S., West, S. G., Sechrest, L., Reno, R., Roediger, H. L., Scarr, S., et al. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734.

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin, 115,* 308–314.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analyses. *Personnel Psychology, 31,* 233–242.

Bartol, K. M., & Manhardt, P. J. (1979). Sex differences in job outcome preferences: Trends among newly hired college graduates. *Journal of Applied Psychology, 64,* 477–482.

Bloom, M., & Milkovich, G. T. (1998). Relationships among risk, incentive pay, and organizational performance. *Academy of Management Journal, 41,* 283–297.

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement, 4,* 385–398.

Bobko, P., & Russell, C. J. (1994). On theory, statistics, and the search for interactions in the organizational sciences. *Journal of Management, 20,* 193–200.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society Series B, 26,* 211–246.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97.

Brock, J. K.-U. (2003). The "power" of international business research. *Journal of International Business Studies, 34,* 90–99.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41,* 25–42.

Chaplin, W. F. (1997). Personality, interactive relations, and applied psychology. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 873–890). San Diego, CA: Academic Press.

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

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Revised Ed.). New York: Academic Press.

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

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

Eden, D. (2002). Replication, meta-analysis, scientific progress, and *AMJ*'s publication policy. *Academy of Management Journal, 45,* 841–846.

Eden, D., & Zuk, Y. (1995). Seasickness as a self-fulfilling prophecy: Raising self-efficacy to boost performance at sea. *Journal of Applied Psychology, 80,* 628–635.

Fichman, M. (1999). Variance explained: Why size does not (always) matter. *Research in Organizational Behavior, 21,* 295–331.

Fisicaro, S. A., & Lautenschlager, G. J. (1992). Power and reliability: The case of homogeneous true score regression across treatments. *Educational and Psychological Measurement, 52,* 505–511.

Ford, J. K., & Noe, R. A. (1987). Self-assessed training needs: The effects of attitudes toward training, management level, and function. *Personnel Psychology, 40,* 25–42.

Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management, 62,* 801–807.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121,* 200–206.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician, 55,* 19–24.

Ibarra, H. (1995). Race, opportunity, and diversity of social circles in managerial networks. *Academy of Management Journal, 38,* 673–703.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics, 18,* 75–86.

Latack, J. C., Josephs, S. L., Roach, B. L., & Levine, M. D. (1987). Carpenter apprentices: Comparison of career transitions for men and women. *Journal of Applied Psychology, 72,* 393–400.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male–female differences: A computer simulation. *American Psychologist, 51,* 157–158.

Mazen, A. M., Graf, L. A., Kellogg, C. E., & Hemmasi, M. (1987). Statistical power in contemporary management research. *Academy of Management Journal, 30,* 369–380.

Mazen, A. M., Hemmasi, M., & Lewis, M. F. (1987). Assessment of statistical power in contemporary strategy research. *Strategic Management Journal, 8,* 403–410.

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49,* 103–120.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286.

Peters, L. H., Fisher, C. D., & O'Connor, E. J. (1982). The moderating effect of situational control of performance variance on the relationship between individual differences and performance. *Personnel Psychology, 35,* 609–621.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27,* 52–71.

Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology, 77,* 336–342.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of source adjustment in preemployment testing. *American Psychologist, 49,* 929–954.

Salgado, J. F. (1998). Sample size in validity studies of personnel selection. *Journal of Occupational and Organizational Psychology, 71,* 161–164.

Schaubroeck, J., May, D. R., & Brown, F. W. (1994). Procedural justice explanations and employee reactions to economic hardship: A field experiment. *Journal of Applied Psychology, 79,* 455–460.

Schmidt, F. L. (1988). The problem of group differences in ability scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt, F. L. (2002). The role of general cognitive ability in job performance: Why there cannot be a debate. *Human Performance, 15,* 187–210.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

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

Society for Industrial and Organizational Psychology, Inc. (1999). *SIOP annual membership directory.* Bowling Green, OH: Author.

Starbuck, B., & Mezias, J. (1996). Journal impact ratings. *The Industrial–Organizational Psychologist, 33(4),* 101–105.

Stone-Romero, E. F., Alliger, G. M., & Aguinis, H. (1994). Type II error problems in the use of moderated multiple regression for the detection of moderating effects of dichotomous variables. *Journal of Management, 20,* 167–178.

Stone-Romero, E. F., & Liakhovitski, D. (2002). Strategies for detecting moderator variables: A review of conceptual and empirical issues. *Research in Personnel and Human Resources Management, 21,* 333–372.

Tepper, B. J. (1994). Investigation of general and program specific attitudes toward corporate drug testing policies. *Journal of Applied Psychology, 79,* 392–401.

Tubbs, M. E. (1993). Commitment as a moderator of the goal–performance relation: A case for clearer construct definition. *Journal of Applied Psychology, 78,* 86–97.

Uhl-Bien, M., & Graen, G. B. (1998). Individual self-management: Analysis of professionals' self-managing activities in functional and cross-functional work teams. *Academy of Management Journal, 41,* 340–350.

Van Dyne, L., & Ang, S. (1998). Organizational citizenship behavior of contingent workers in Singapore. *Academy of Management Journal, 41,* 692–703.

Vecchio, R. P. (1980). A test of a moderator of the job satisfaction–job quality relationship: The case of religious affiliation. *Journal of Applied Psychology, 65,* 195–201.

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

Wigdor, A. K., & Garner W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies.* (Report of the National Research Council Committee on Ability Testing). Washington, DC: National Academy of Sciences Press.

Williams, K. J., Suls, J., Alliger, G. M., & Learner, S. M. (1991). Multiple role juggling and daily mood states in working mothers: An experience sampling study. *Journal of Applied Psychology, 76,* 664–674.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin, 76,* 295–310.

Appendix A

Computation of Effect Size: Modified f^2

The hypothesis of no moderating effect in MMR is tested by comparing $F_m = \dfrac{SSI/(k-1)}{SSE/(N-2k)}$ to $F^{1-\alpha}_{k-1,N-2k}$, where k is the number of moderator-based subpopulations, N is the total sample size (across all groups), SSI is the sum of squares due to the interaction between the categorical moderator variable Z and the continuous predictor X (cf. Equation 1), and SSE is the error sum of squares after fitting first-order effects and product terms. It can be shown that conditional on X,

$$\mathrm{E}(SSI) = \sum_{i=1}^{k} \sigma_i^2(1-\rho_i^2)(1-w_i) + \sum_{i=1}^{k}(n_i-1)\rho_i^2\sigma_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i-1)\rho_i\sigma_i s_{x_i}\right]^2}{\sum_{i=1}^{k}(n_i-1)s_{x_i}^2} \qquad \text{(A1)}$$

and

$$\mathrm{E}(SSE) = \sum_{i=1}^{k}(n_i-2)\sigma_i^2(1-\rho_i^2),$$

where $w_i = \dfrac{(n_i-1)s_{x_i}^2}{\sum_{j=1}^{k}(n_j-1)s_{x_j}^2}$, σ_i^2 is the variance of the observable Y scores in the i^{th} subpopulation, $s_{x_i}^2$ is the sample variance of X in the i^{th} subpopulation, n_i is the sample size from the i^{th} subpopulation, and ρ_i is the correlation between Y and X in the i^{th} subpopulation.

The first term in E(SSI) reflects the $k - 1$ degrees of freedom associated with SSI. The remaining terms in E(SSI) reflect variation accounted for by the categorical moderator times continuous predictor variable interaction. Accordingly, the effect size is

$$f^2 = \frac{\mathrm{E}(SSI) - \sum_{i=1}^{k}\sigma_i^2(1-\rho_i^2)(1-w_i)}{\mathrm{E}(SSE)}$$

$$= \frac{\sum_{i=1}^{k}(n_i-1)\rho_i^2\sigma_i^2 - \dfrac{\left[\sum_{i=1}^{k}(n_i-1)\rho_i\sigma_i s_{x_i}\right]^2}{\sum_{i=1}^{k}(n_i-1)s_{x_i}^2}}{\sum_{i=1}^{k}(n_i-2)\sigma_i^2(1-\rho_i^2)}.$$

If ρ_i is written as $\rho_i = \beta_i s_{x_i}/\sigma_i$ then effect size can be written as

$$f^2 = \frac{\sum_{i=1}^{k} w_i(\beta_i-\bar{\beta})^2}{\theta - \bar{\beta}^2 - \sum_{i=1}^{k} w_i(\beta_i-\bar{\beta})^2} + O\!\left(\frac{1}{N}\right),$$

where

$$\theta = \frac{\sum_{i=1}^{k}(n_i-1)\sigma^2}{\sum_{i=1}^{k}(n_i-1)s_{x_i}^2},$$

$$\bar{\beta} = \sum_{i=1}^{k} w_i\beta_i,$$

and w_i is defined in Equation A1. Note that if $\beta_i = \rho_i\sigma_i/s_{x_i}$ is constant for all i (i.e., $\beta_1 = \bar{\beta}$), then $f^2 = 0$. To estimate effect size, sample quantities can be substituted for population parameters.

To examine power for a selected (i.e., hypothetical) effect size, say f^2_{sel}, the quantities n_i, σ_i^2, and $s_{x_i}^2$ for $i = 1, \ldots, k$ are held fixed at their observed values and new values β^*_{i} for $i = 1, \ldots, k$ are obtained to satisfy

$$f^2_{\text{sel}} = \frac{\sum_{i=1}^{k} w_i(\beta^*_i - \bar{\beta}^*)^2}{\theta - \bar{\beta}^{*2} - \sum_{i=1}^{k} w_i(\beta^*_i - \bar{\beta}^*)^2}. \qquad \text{(A2)}$$

In this article β^*_i was computed as

$$\beta^*_i = \beta_i + \gamma_i(\beta_i - \bar{\beta}) \text{ for } i = 1, \ldots, k. \qquad \text{(A3)}$$

The multipliers $\gamma_i, \ldots, \gamma_k$ were selected to minimize the quantity

$$\sum_{i=1}^{k}(\gamma_i - \bar{\gamma})^2,$$

where $\bar{\gamma} = \frac{1}{k}\Sigma_{i=1}^{k}\gamma_i$, subject to the constraints (a) Equation A2 is satisfied and (b) $\rho^*_{\text{i}} = \beta^*_{\text{i}} s_{x_i}/\sigma_i \in [-0.99, 0.99]$ for each i. In many cases, the solutions for $\gamma_1, \ldots, \gamma_k$ are

$$\gamma_i = \gamma \text{ for all } i, \text{ where } \gamma = \left(\frac{f^2_{sel}(\theta - \bar{\beta}^2)}{(1+f^2_{sel})\sum_{i=1}^{k} w_i(\beta_i-\bar{\beta})^2}\right)^{\frac{1}{2}} - 1.$$

If f^2 is substantially smaller than f^2_{sel}, then the above solution may not satisfy $\beta^*_i s_{x_i}/\sigma_i \in [-0.99, 0.99]$ for all i. In this case, the multiples $\gamma_1, \ldots, \gamma_k$ cannot be chosen to be identical.

If the multipliers $\gamma_1, \ldots, \gamma_k$ can be chosen to be identical, then the modification merely adjusts the magnitude of the observed differences among the regression coefficients but retains the relative differences. Alternative procedures for computing β^*_i could be devised. The procedure in Equation A3 was chosen because it retains the observed pattern of regression coefficients to the largest extent possible.

Power corresponding to the selected effect f^2_{sel} was computed using the algorithm described in Aguinis et al. (2001) and shown in Appendix C.

(*Appendixes continue*)

Appendix B

Computation of Construct-Level Effect Sizes

The question to be addressed is the following: What values do effect sizes take on if X and Y reliability are increased from original values to 1.0? Stated differently, the construct-level effect sizes show the magnitude of the moderating effect when X and Y reliability are increased from their original values (i.e., the value reported in the original article if available or .80) to 1.0.

Denote the correlation between the continuous predictor X and the continuous criterion Y in a subpopulation (i.e., moderator-based category) as ρ_{true} and denote the correlation between the observable scores as $\rho_{\text{observable}}$. If original X and Y reliabilities are α_x and α_y, respectively, then $\rho_{\text{observable}} = \rho_{\text{true}}\sqrt{\alpha_x\alpha_y}$. The value of ρ_{true} was computed as

$$\rho_{\text{true}} = \begin{cases} \min\left(\dfrac{\rho_{\text{observable}}}{\sqrt{\alpha_x\alpha_y}},\ 0.999\right) & \text{if } \rho_{\text{observable}} > 0 \\ \max\left(\dfrac{\rho_{\text{observable}}}{\sqrt{\alpha_x\alpha_y}},\ -0.999\right) & \text{if } \rho_{\text{observable}} < 0 \\ 0 & \text{if } \rho_{\text{observable}} = 0. \end{cases}$$

After the value of ρ_{true} was computed, consider that the X and Y reliabilities are now equal to 1, rather than their original values (or 0.80). Accordingly, $\rho_{\text{observable}}$ is now equal to ρ_{true}. The new value of $\rho_{\text{observable}}$, therefore, is larger than the original value of $\rho_{\text{observable}}$. The new value of $\rho_{\text{observable}}$ can be interpreted as the observable correlation that would exist if the study could be replicated, but with X and Y having perfect reliability. The new values of $\rho_{\text{observable}}$ are entered directly into the equations in Appendix A to obtain the effect sizes that would exist if the study could be replicated, but with X and Y having perfect reliability.

Appendix C

Power Approximation (from Aguinis, Boik, & Pierce, 2001, pp. 319–320)

The power of the MMR F test is

Power

$$\approx \Pr\left[\left(\frac{k-1}{N-2k}\right)F^{1-\alpha}_{k-1,N-2k}\sum_{j=1}^{k}\frac{\sigma^2_{y,j}(1-\rho_j^2\alpha_{x,j}\alpha_{y,j})}{\alpha_{y,j}}H_j - \sum_{j=1}^{k-1}\omega_j G_j \leq 0\right],$$

where k is the number of moderator-based subpopulations, $\sigma^2_{y,j}$ is the variance of the true Y scores in subpopulation j, ρ_j^2 is the squared correlation between the true X and Y scores in subpopulation j, $\alpha_{x,j}$ is the reliability for X in subpopulation j, $\alpha_{y,j}$ is the reliability for Y in subpopulation j, and ω_j is the j^{th} eigen-value of $(\mathbf{C}'\mathbf{D}\mathbf{C})^{-1}\ \mathbf{C}'\mathbf{V}\mathbf{C}$;

$$\mathbf{D} = \text{Diag}\left[\frac{\alpha_{x,j}(n_j+1)}{(n_j-1)^2\delta_j\sigma^2_{x,j}};\ j = 1, \ldots, k\right];$$

$$\mathbf{V} = \text{Diag}\left[\frac{\sigma^2_{y,j}\alpha_{x,j}(1-\rho_j^2\alpha_{x,j}\alpha_{y,j})(n_j2+1)}{\alpha_{y,j}(n_j-1)^2\delta_j\sigma^2_{x,j}};j = 1, \ldots, k\right];$$

where n_j is the size of subpopulation j, δ_j is the ratio of the expected sample variance of X to the population variance of X in subpopulation j, $\sigma^2_{x,j}$ is the variance of the true X scores in subpopulation j, 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 λ_j is a noncentrality parameter;

$$\lambda_j = \frac{(\mathbf{u}_j'\mathbf{C}'\beta_1)^2}{2\mathbf{u}_j'\mathbf{C}'\mathbf{V}\mathbf{C}\mathbf{u}_j};$$

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

Note. This appendix is adapted from H. Aguinis, R. J. Boik, & C. A. Pierce, *Organizational Research Methods, 4,* pp. 291–323, copyright © 2001 by Sage Publications. Reprinted by Permission of Sage Publications, Inc.

Appendix D

Technical Note on Computation of f^2 for Given Statistical Power and Sample Size

This note uses the same notation as that in Appendix A. Using the results of Khatri (1966), it can be shown that

$$\mathbf{C}(\mathbf{C}'\mathbf{D}_x\mathbf{C})^{-1}\mathbf{C} = \mathbf{D}_x^{-1} - \mathbf{D}_x^{-1}\mathbf{1}_k(\mathbf{1}_k'\mathbf{D}_x^{-1}\mathbf{1}_k)^{-1}\mathbf{D}_x^{-1}.$$

If $\sigma_i^2 = \sigma^2$ for all i, $\alpha_{x_i} = \alpha_{y_i} = 1$ for all i, $n_i = n$ for all i, and $S^2_{x_i} = S^2_x$ for all i, then the above result can be used to show that the expected value of the numerator quadratic form in the F statistic is

$$\text{E}(SSI) = \text{E}[\hat{\beta}_1'\mathbf{C}(\mathbf{C}'\mathbf{D}_x\mathbf{C})^{-1}\mathbf{C}\hat{\beta}_1]$$

$$= \frac{1}{k}(k-1)\sigma^2\sum_{i=1}^{k}(1-\rho_i^2) + \frac{1}{k}S_{xx}\sum_{i=1}^{k}(\beta_i - \bar{\beta})^2,$$

where $S_{xx} = (n-2)S_x^2$ and β_i is the i^{th} component of β_1. Under these same assumptions, the expected value of the denominator quadratic form of the F statistic is

$$\text{E}(SSE) = (n-2k)\frac{1}{k}\sigma^2\sum_{i=1}^{k}(1-\rho_i^2).$$

Dividing the numerator and denominator quadratic forms by $\frac{1}{k}\sigma^2\sum_{i=1}^{k}(1 - \rho_i^2)$ yields

$$\frac{E(SSI)}{\frac{1}{k}\sigma^2\sum_{i=1}^{k}(1-\rho_i^2)} = k-1+\frac{S_{xx}\sum_{i=1}^{k}(\beta_i-\bar{\beta})^2}{\sigma^2\sum_{i=1}^{k}(1-\rho_i^2)}$$

and

$$\frac{E(SSE)}{\frac{1}{k}\sigma^2\sum_{i=1}^{k}(1-\rho_i^2)} = N-2k.$$

Note that these moments are identical to the moments of a noncentral F random variable with $k-1$ numerator degrees of freedom, $N-2k$ denominator degrees of freedom, and noncentrality parameter

$$\lambda = \frac{S_{xx}\sum_{i=1}^{k}(\beta_i-\bar{\beta})^2}{2\sigma^2\sum_{i=1}^{k}(1-\rho_i^2)}.$$

Accordingly, if $\sigma_i^2 = \sigma^2$ for all *i*, $\alpha_{x_i} = \alpha_{y_i} = 1$ for all *i*, $n_i = n$ for all *i*, and $S_{x_i}^2 = S_x^2$ for all *i*, and if normality is satisfied, then the MMR F statistic is distributed as $F_{k-1,N-2k,\lambda}$ to a first-order approximation. Furthermore, under these assumptions, the effect size is

$$f^2 = \frac{S_{xx}\sum_{i=1}^{k}(\beta_i-\bar{\beta})^2}{(N-2k)\sigma^2\sum_{i=1}^{k}(1-\rho_i^2)} = \frac{2\lambda}{N-2k}.$$

To determine the value of f^2 that corresponds to a specific power, the following equation must be solved for λ:

$$P(F_{k-1,N-2k,\lambda} > F^{1-\alpha}_{k-1,N-2k}) = \text{power},$$

where $F^{1-\alpha}_{k-1,N-2k}$ is the 100 $(1-\alpha)$ percentile of the central F distribution with $k-1$ and $N-2k$ degrees of freedom. The solution to the above equation cannot be written in closed form. The equation must be solved numerically. After solving for λ, the associated f^2 is

$$f^2 \approx \frac{2\lambda}{N-2k},$$

which is Equation 2 shown in text.

Received June 12, 2002
Revision received November 20, 2003
Accepted November 25, 2003 ■