Journal of Applied Psychology
2006, Vol. 91, No. 1, 1–5

0021-9010/06/$12.00 DOI: 10.1037/0021-9010.91.1.1

Statistical Power and Parameter Stability When Subjects Are Few and Tests Are Many: Comment on Peterson, Smith, Martorana, and Owens (2003)

John R. Hollenbeck, D. Scott DeRue, and Michael Mannor
Michigan State University

This comment illustrates how small sample sizes, when combined with many statistical tests, can generate unstable parameter estimates and invalid inferences. Although statistical power for 1 test in a small-sample context is too low, the experimentwise power is often high when many tests are conducted, thus leading to Type I errors that will not replicate when retested. This comment's results show how radically the specific conclusions and inferences in R. S. Peterson, D. B. Smith, P. V. Martorana, and P. D. Owens's (2003) study changed with the inclusion or exclusion of 1 data point. When a more appropriate experimentwise statistical test was applied, the instability in the inferences was eliminated, but all the inferences become nonsignificant, thus changing the positive conclusions.

Keywords: power, sample size, teams

In terms of research methodology, perhaps one of the most significant advancements over the past 25 years has been the development and refinement of meta-analytic techniques for establishing parameter estimates, testing moderators, and drawing inferences about relationships (Cohen, 1994; Rosnow & Rosenthal, 1991). As Hunter and Schmidt (2004) have argued, prior to the advent of meta-analysis, most researchers did not appreciate the impact that sampling error had on the stability of effect sizes. Moreover, when the process of generating unstable parameters is combined with the use of statistical hypothesis inference testing for drawing dichotomous conclusions about the presence or absence of relationships, the opportunity to generate spurious inferences that fail to replicate at a later time is excessive (Schmidt, 1996). Time and time again over the past 25 years, meta-analytic researchers have revisited and reversed past conclusions from narrative reviews of literatures that were plagued by sampling errors and other statistical artifacts.

Although one might think that the problems associated with small sample sizes are common knowledge, evidence exists to suggest otherwise. For instance, Mone, Mueller, and Mauland (1996) documented that both the applied psychological and management literatures consistently use low levels of statistical power and that the level of power in these literatures has not changed over time—all despite numerous calls for formal power analyses as a basic requirement in the methods sections of studies that apply quantitative analyses (Cohen, 1962, 1977, 1988, 1992). Although low statistical power can result from a variety of factors that produce a reduced sample-based effect size vis-à-vis the population counterpart (e.g., measurement error and range restriction), small sample size has been and continues to be the primary cause of low statistical power in applied research.

As Maxwell (2004) recently noted, one primary reason for the continued use of low power studies is that although the power for any one test with a small sample size is low, the experimentwise statistical power (i.e., the ability to find something, anything) is quite high when one conducts a large number of nonindependent tests in one setting, as is often the case in applied research. Thus, it is possible to obtain significant results at the study level even when the sample size is very small at the parameter level. The problem, of course, is that each of these parameter-level effects is impacted tremendously by sampling error, is highly unstable, and therefore is almost impossible to replicate. Thus, Maxwell (2004) concluded "small sample sizes lead to a published research literature that is virtually guaranteed to contain numerous inconsistencies about what is statistically significant and what is not" (p. 155).

A recent study by Peterson, Smith, Martorana, and Owens (2003), in our opinion, displays the dangers associated with the practice of combining small sample sizes with a large number of statistical tests. The purpose of our comment is to show how low statistical power as a result of small sample size leads to the drawing of inferences that are not robust and are much too sensitive to sampling error. Although similar calls for greater appreciation of statistical power have gone unheeded (Mone et al., 1996), the study reported here uses a graphic and intuitive simulation to show the extent to which low statistical power can lead to unstable and/or invalid inferences, and hence it may have a broader impact in communicating the salience of this issue to applied researchers.

From the outset, we should note that Peterson et al.'s (2003) study was a highly creative approach to a very important problem. The Q sort method, although not a traditional form of qualitative research (e.g., ethnography, semiotic analysis, dramaturgical analysis, deconstruction), is an interesting technique for generating quantitative data. Furthermore, the theoretical model that they

John R. Hollenbeck, D. Scott DeRue, and Michael Mannor, Department of Management, Michigan State University.

Correspondence concerning this article should be addressed to John R. Hollenbeck, Department of Management, Michigan State University, East Lansing, MI 48824-1117. E-mail: jrh@bus.msu.edu

developed—linking chief executive officer (CEO) personality traits to organizational performance through top management team (TMT) dynamics—is both credible and in line with existing evidence. Peterson et al. also were very open in publicly sharing most of their data in the article itself. Finally, they were very frank about the limitations of their study, listing five specific cautions: the narrow nature of the sample (all large U.S. firms), the second-hand nature of the data, potential performance-related biases that informed raters' judgments, potential historical inaccuracies in the popular press reports of team dynamics, and finally, the inability to test for nonlinear relationships.

However, they did not identify how the small size of the sample, combined with repeated use of statistical hypothesis inference testing, created a set of results and corresponding inferences that were extremely unstable. In the study reported here, we reanalyzed Peterson et al.'s (2003) data and show how radically the inferences change depending on the inclusion or exclusion of single data points. The results of our study qualify many of Peterson et al.'s specific conclusions about CEO personalities, team dynamics, and firm performance, but in a larger sense, our results also speak to the dangers of perpetuating research that compromises statistical power at the parameter level. In the following sections, we briefly review the design of Peterson et al.'s study, the nature of the data they collected, their analysis, and the inferences they drew. We then show how sensitive these inferences are to sampling error by demonstrating how the inclusion or exclusion of a single data point changes the inferences.

Research Design and Sample Size Limitations

In their study of CEO personality, team dynamics, and firm performance, Peterson et al. (2003) separated the CEO from each firm's TMT, enabling study of specific CEO characteristics and the effects of these characteristics on TMT dynamics and ultimately on organizational performance operationalized as income growth. On the basis of the prior literature, they explicitly stated 11 directional hypotheses dealing with relationships between CEO traits and dimensions of team dynamics and 4 expected relationships between team dynamics and firm performance (see Peterson et al., 2003, p. 803, the underlined values in Tables 3 and 4). In addition to these 15 expected relationships, Peterson et al. explored all of the remaining 33 possible correlations between traits, team dynamics, and firm performance, and they subsequently discussed those that reached statistically significant levels. Thus, in the end, they tested 48 different correlations for statistical significance.

Unlike research with line workers, first-line supervisors, salespeople, nurses, military recruits, or college sophomores, it is difficult to get a large number of CEOs to participate in research. Hence, the 48 correlations described in the previous paragraph were computed on the basis of a sample of 17 people (see Peterson et al., 2003, Table 1, p. 801). The CEOs did not provide the data themselves; instead, the researchers had raters evaluate archival data taken from books and popular press reports that described the CEOs or their teams. Each rater read through the qualitative information and then used a structured questionnaire to answer questions about the problem of interest.

This is a very innovative method to derive quantitative data from nontraditional sources, such as the popular press literature. However, Peterson et al. (2003) unfortunately chose to limit their study to a sample of 17 CEOs. This is an understandable decision given the labor-intensive nature of this method. Also, this sample size is comparable with what one might normally encounter in studies using traditional qualitative methods, such as content analysis or ethnography. Furthermore, this sample size may be completely acceptable in exploratory studies. Small sample sizes should not and do not limit researchers' ability to conduct exploratory research and search for insights that would be interesting to explore with other techniques and larger sample sizes. However, the small sample size combined with the decision to use statistical hypothesis inference testing (a confirmatory rather than exploratory procedure) ultimately generates a set of parameter estimates that are highly unstable and inferences that are not very robust.

This unstable evidentiary base places future researchers wishing to build from this theory, or more important, practitioners thinking about implementing it through executive selection, on a very weak foundation. That is, on the basis of the results presented in Peterson et al.'s (2003) study, someone in the area of executive selection might use this evidence to justify selecting or rejecting candidates on the basis of scores on some of the dimensions examined in the research. However, the parameter estimates that are calculated from this sample are highly unstable, and below we discuss how simply removing 1 subject from the sample changes the inferences reported by Peterson et al.

Reanalysis of Peterson et al.'s (2003) Data

In Peterson et al.'s (2003) Tables 3 and 4, they underlined the 15 correlations that represent tests of hypotheses or formal expected relationships. In 13 of the 15 cases, these relationships were supported with a one-tailed test at the .10 level of statistical significance. Furthermore, in 8 cases, statistically significant results were obtained in the absence of a formal hypothesis or expectation. Although these correlations were not attached to formal hypotheses, most of them were tested with directional one-tailed tests at the .10 level, and many were mentioned in their discussion section. Such broad support suggests that the model they were proposing is quite valid and can serve as a foundation on which to build future research and practice.

The 17 inferences about the relationship between CEO personality and TMT dynamics that one would draw from these significance tests are described in Table 1 of this comment. However, given the small sample size on which these tests were based, the broad support found in this study depends largely on the exact and precise specification of the null hypothesis and selected sample. In other words, the exact 17 CEOs used in the analysis define the results, and the inferences can be supported only by using a highly liberal one-tailed .10 probability level. To illustrate the sensitivity of these results and inferences to the specific sample and probability level, we reanalyzed the data from Peterson et al.'s (2003) study. The objective of this reanalysis was to assess the degree to which their conclusions would change by simply removing 1 subject from the sample or by using a more traditional probability level.

On the basis of the same sample of 17 CEOs (see Table 1 from Peterson et al., 2003), we first reexamined all of the relationships found to be significant in the original study. Each time the simu-

Table 1
Simulation Results

Hypothesized and post hoc inferences	Is the inference significant? Peterson et al. (2003)	If Ron Miller is excluded	If Roberto Goizueta is excluded	Total
Neuroticism–rigidity[a]	yes	maybe	maybe	2
Neuroticism–leader weakness[a]	yes	no	yes	1
Neuroticism–factionalism[a]	yes	yes	maybe	13
Neuroticism–legalism	yes	yes	maybe	12
Extraversion–leader weakness[a]	yes	no	yes	1
Openness–rigidity[a]	maybe	no	no	17
Openness–leader weakness	yes	no	no	4
Openness–risk aversion[a]	yes	no	no	10
Agreeableness–factionalism[a]	yes	yes	yes	1
Agreeableness–legalism	yes	yes	yes	1
Agreeableness–decentralization[a]	yes	yes	yes	1
Conscientiousness–rigidity	yes	maybe	yes	1
Conscientiousness–control[a]	maybe	maybe	no	6
Conscientiousness–leader weakness	maybe	no	no	4
Conscientiousness–factionalism	yes	yes	maybe	3
Conscientiousness–legalism[a]	yes	yes	yes	0
Conscientiousness–decentralization	yes	yes	yes	1

Note. *Yes* indicates that the inference is supported at $p < .05$, one tailed. *Maybe* indicates that the inference is supported at $p < .10$, one-tailed (but not at $p < .05$, one-tailed). *No* indicates that the inference is not supported at $p < .10$, one-tailed. Each number in the *Total* column represents how many times, in an absolute sense, the specific inference changed after the exclusion of a single chief executive officer (CEO). A value of 0 implies that the result was robust to the removal of any one CEO and did not change. A value of 17 implies that the result was totally unstable and changed when any one of the CEOs was removed from the sample.
[a] This relationship was an a priori hypothesis from Peterson et al.'s (2003) study.

lation was run, a different CEO was excluded from the sample. Thus, 17 different tests with a sample size of 16 subjects were performed as part of the simulation. The simulation was initially conducted using the same one-tailed significance level ($p < .10$) used in Peterson et al.'s (2003) article; subsequently, we also performed the simulation using the more traditional and conservative approach ($p < .05$).

As a specific example, our Table 1 provides the inferences that one would have drawn from these data had Disney's Ron Miller been eliminated from the data set. The resulting new set of inferences is, in fact, substantially different from that originally published. In the original study, Peterson et al. (2003) found that TMT leader weakness–dominance characteristics are affected by the 4 CEO characteristics of neuroticism, extraversion, openness, and conscientiousness (see their Table 1). With just Ron Miller excluded from the data, we can no longer support the conclusion that any of these 4 CEO personality traits significantly impact TMT leader weakness–dominance. In addition, the original study found that TMT rigidity–flexibility was affected by the 3 CEO personality traits of neuroticism, openness, and conscientiousness. With just Ron Miller excluded from the data, all three conclusions about the impact of these CEO personality traits on TMT rigidity–flexibility become questionable. In fact, with the removal of Ron Miller from the sample, 9 of the 17 inferences about how CEO personality traits affect TMT dynamics that were originally found to be significant become nonsignificant on the basis of statistical hypothesis inference testing.

As another example, our Table 1 provides the inferences that one would have drawn from these data if Ron Miller were reinserted and instead Roberto Goizueta (Coca-Cola) were removed from the sample. In this scenario, some of the inferences that were lost when Ron Miller was excluded reappear, but other inferences that were originally supported now become questionable. For instance, the relationship between CEO neuroticism and TMT leader weakness–dominance reappears as significant, but the link between CEO neuroticism and TMT factionalism–cohesion now becomes questionable. Similarly, after removing Roberto Goizueta, the inferences that were originally established between (a) CEO openness and TMT risk aversion–risk taking behavior and (b) CEO conscientiousness and TMT control–crisis characteristics are no longer supported.

Not only do inferences supported in the original study become disputed as a function of the simulation, but new inferences not made in the original study can also be discovered. In one case, when Kodak's Kay Whitmore is the only CEO excluded from the sample, a new conclusion can be reached—CEO agreeableness has a significant impact on TMT rigidity–flexibility. It is interesting that this relationship was not found in any of the previous analyses.

To summarize the results of the simulation, the last column in our Table 1 shows how many times, in an absolute sense, the specific inferences change after the exclusion of a single CEO. A value of 0 implies that the result is robust to the removal of any 1 CEO and does not change. A value of 17 implies that the result is totally unstable and changes when any 1 of the CEOs is removed from the sample. Only 1 of the 17 relationships between CEO personality and TMT dynamics was robust to the exclusion of any single CEO, whereas some of the links were so tenuous that the removal of any 1 of the 17 CEOs studied resulted in the inference not being supported.

One solution to the stability problem inherent in Peterson et al.'s (2003) study would be to use an adjustment procedure that controls the experimentwise Type I error rate (see also Maxwell, 2004).[1] For example, the Bonferroni correction would have used a .05/48 ($p = .001$) a priori alpha level for the significance tests. If Peterson et al.'s original analysis had been conducted using a two-tailed .001 level as the threshold for statistical significance, this would indeed have eliminated all the instability that we note in our Table 1. That is, the inferences based on those data would not change after the removal of any single data point. Unfortunately, none of the correlations that were significant in the original study would remain significant, and thus stability is achieved by inferring that all 48 of the relationships are not significant. This differs markedly from the conclusion reached by Peterson et al. that "our results provided broad support for our general hypothesis that CEO personality affects TMT group dynamics and that TMT group dynamics are related to organizational performance" (p. 802).

As is apparent from our Table 1, replicating Peterson et al.'s (2003) results is extremely difficult, even when one merely removes a single subject but retains the rest of the original sample and criterion.[2] Clearly, anyone trying to replicate these findings with a different sample of CEOs or a different set of criteria would experience a serious, and in all likelihood, insurmountable challenge. With respect to different criteria, Peterson et al. noted,

> We collected several measures of organizational performance (including growth in sales, return on investment, and return on assets) that were all highly correlated in this study, as well. Indeed, we obtain the same pattern of results if we use income growth, sales growth, or change in return on investment. (p. 801)

However, with a sample size this small, it would be remarkable if the pattern of results would not be sensitive to different operationalizations of firm performance. To test this, we conducted an independent test of the relationship between TMT dynamics (using the data provided by Peterson et al., 2003) and organizational performance using five alternative measures (return on assets, return on equity, and three forms of income growth: net income growth, operating income growth before depreciation, and operating income growth after depreciation). All of these measures of organizational performance are widely used in the strategy literature (Krishnan, Miller, & Judge, 1997; McGuire, Sundgren, & Schneeweis, 1988; Nehrt, 1996) with return on assets and return on equity being the most common (Dalton & Kesner, 1985; Iaquinto & Fredrickson, 1997). The data for organizational performance were obtained from the COMPUSTAT database.

Using these five different measures, we were unable to find consistent support for any of the four relationships between TMT dynamics and organizational performance that were suggested by Peterson et al. (2003). In fact, some of the relationships supported in the original study are not even directionally consistent with the results of our analysis. In one example, where Peterson et al. found that higher TMT optimism orientation led to negative income growth, three of our five correlations for this factor actually indicated a positive relationship with organizational performance. If the authors had used return on equity instead of income growth, the relationships between TMT factionalism and risk aversion with organizational performance would not have been supported. Furthermore, three completely new relationships between TMT dynamics (TMT control, legalism, decentralization) and organizational performance would have been discovered. Thus, the decision to use the specific measure of income growth (as calculated in the original study) does appear to be important in determining the results because, as one might suspect, they are not robust to alternative operationalizations of firm performance. We welcome other researchers to test these relationships using other alternative organizational performance measures or a different set of CEOs. However, we suspect that instability will be the rule, not the exception.

Discussion

The difficulty in conducting quality research on CEOs and TMTs should not discourage the future pursuit of such endeavors. In fact, the research done by Peterson et al. (2003) provides a solid theoretical base and a plausible process model that may ultimately serve as a useful vehicle for understanding CEOs, their TMTs, and their firms. Notably, Peterson et al. showed how the Q sort method can be used to empirically examine content published in the popular press. In all the work done on organizations, data on CEOs may be the most difficult to collect. The extensive access that the popular press provides to these exclusive individuals offers an abundant wealth of data. Our reservations expressed about Peterson et al.'s study should not reflect poorly on the TMT research domain or the Q sort methodology, which we feel is a very creative approach that may serve as a model for others to build on.

Instead, our major concern is with the process of using statistical hypothesis inference testing in a context where there are many tests but few subjects. Whereas a richly described set of qualitative inductions from this data set might have made a highly valuable contribution to this literature—as would a large-sample deductive test of specific a priori hypotheses—the quantitative approach to this particular data set, given its small sample size, creates some major difficulties when it comes to drawing rigorous inferences.

Clearly, it is important to point out that both qualitative and quantitative approaches can and have been used effectively in the study of executives, team dynamics, and organizations when matched with an appropriate sample. For instance, Ashcraft (1999) studied leader succession in executives using a well-designed study based on ethnography (a qualitative technique). Other examples include studies by Morris and Moore (2000) and Jehn (1997), who used other qualitative techniques to effectively study team dynamics and organizations. Some studies, such as those by Wageman (1995) and Edmondson (1999), effectively integrated both qualitative and quantitative techniques. Peterson et al.'s (2003) study, on the other hand, takes a quantitative approach to data that are largely qualitative, but does not

[1] We thank an anonymous reviewer for suggesting the possibility that Peterson et al. (2003) could have used an adjustment procedure that controls the experimentwise Type I error rate.

[2] The simulation presented in this article was conducted with the original data set ($N = 17$) used in Peterson et al.'s (2003) study. Actual results from this simulation are therefore subject to the same methodological inadequacies (low statistical power, parameter instability) as were in the original study and thus should not be interpreted as substantive findings. The purpose of these analyses was simply to document the instability of the original data.

supplement this with any of these other, more formal qualitative techniques. Indeed, even if one stays within the quantitative domain, there are other approaches to research design (such as using within-subjects designs with repeated measures or optimal research designs that use selective sampling to increase the variance in predictors and criteria) that might be used to help increase statistical power given a fixed sample size.

In fact, instead of relying on alternative research design features or more traditional qualitative techniques for analyzing these data, Peterson et al. (2003) actually leaned even further in the quantitative direction, and at one point, they performed a canonical correlational analysis. For example, they reported the results of a canonical correlation where there were 5 predictors, 8 criterion, and only 18 subjects (Peterson et al., 2003, p. 801). Multivariate optimization data analytic techniques such as canonical correlation are even more sensitive to sampling error than are simpler bivariate approaches, and the 13:18 variable-to-subject ratio on which this analysis was based created a set of results that are not likely to stand up well to cross-validation.

It should also be noted that our simulation focused mainly on the inferences drawn about the first component of Peterson et al.'s (2003) model, where they proposed links between CEO personality and TMT dynamics. The original article did not include the data on firm performance, and thus we were not able to directly simulate what would happen to the inferences drawn in the second part of their model, where they proposed links between TMT dynamics and firm performance. Nonetheless, as we showed above, the results associated with organizational performance are highly sensitive to the exact criterion used and do not replicate if one uses any of the five traditional operationalizations of firm performance used in the strategy literature.

Some may argue that the best way to proceed in developing a line of research is to show the way by publishing a small-sample study, with an eye to encouraging later large-scale quantitative replication and assessment of the robustness of effects through meta-analysis. That is, trade-off small samples early in a program of research in the hope that future researchers will replicate results with larger samples. However, we believe that using small sample sizes and statistical hypothesis inference testing to draw dichotomous inferences with respect to a large number of unstable parameters has no countervailing positive outcome against which to trade.

In contrast, we believe that use of this paradigm (i.e., small-sample studies that use statistical hypothesis inference testing to draw a large number of inferences) merely legitimizes more use of this paradigm. Moreover, the subsequent studies using the same paradigm rarely, if ever, replicate the earlier studies with an eye toward refining parameter estimates. Rather, these future studies uncritically accept the results of the earlier study and then go on to extend the findings to new variables using the same paradigm.

In the effort to build a sound and replicable scientific foundation for the study of team dynamics, researchers will need to use many different approaches. Both quantitative and qualitative techniques are fully warranted and necessary for advancing the scientific body of knowledge in this particular research domain. However, as the variety of research techniques used to study team-based phenomena expands, we must fully commit to the rigors of both qualitative and quantitative inquiry and not simply go halfway in our efforts to remake the methods by which we generate and analyze data.

References

Ashcraft, K. L. (1999). Managing maternity leave: A qualitative analysis of temporary executive succession. *Administrative Science Quarterly, 44,* 240–280.

Cohen, J. (1962). The statistical power of abnormal–social psychological research. *Journal of Abnormal and Social Psychology, 65,* 143–153.

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

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

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

Dalton, D. R., & Kesner, I. F. (1985). Organizational performance as an antecedent of inside/outside chief executive succession: An empirical assessment. *Academy of Management Journal, 28,* 749–762.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44,* 350–383.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Iaquinto, A. L., & Fredrickson, J. W. (1997). Top management team agreement about the strategic decision process: A test of some of its determinants and consequences. *Strategic Management Journal, 18,* 63–75.

Jehn, K. A. (1997). A qualitative analysis of conflict types and dimensions in organizational groups. *Administrative Science Quarterly, 42,* 530–557.

Krishnan, H. A., Miller, A., & Judge, W. Q. (1997). Diversification and top management team complementarity: Is performance improved by merging similar or dissimilar teams? *Strategic Management Journal, 18,* 361–374.

Maxwell, S. E. (2004). The persistence of unpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

McGuire, J. B., Sundgren, A., & Schneeweis, T. (1988). Corporate social responsibility and firm financial performance. *Academy of Management Journal, 31,* 854–872.

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

Morris, M. W., & Moore, P. C. (2000). The lessons we (don't) learn: Counterfactual thinking and organizational accountability after a close call. *Administrative Science Quarterly, 45,* 737–765.

Nehrt, C. (1996). Timing and intensity effects of environmental investments. *Strategic Management Journal, 17,* 535–547.

Peterson, R. S., Smith, D. B., Martorana, P. V., & Owens, P. D. (2003). The impact of chief executive officer personality on top management team dynamics: One mechanism by which leadership affects organizational performance. *Journal of Applied Psychology, 88,* 795–808.

Rosnow, R. L., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *Psychological Bulletin, 110,* 574–576.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Wageman, R. (1995). Interdependence and group effectiveness. *Administrative Science Quarterly, 40,* 145–180.

Received March 11, 2004
Revision received October 7, 2004
Accepted October 27, 2004 ■