Journal of Management and Governance **8:** 337–343, 2004.

Views and Debates

Methodological Challenges Posed by Measures of Performance★

WILLIAM H. STARBUCK
ITT Professor of Creative Management, New York University, 40 West Fourth Street, New York, N.Y. 10012-1118, USA (E-mail: wstarbuc@stern.nyu.edu)

Abstract. Performance measures are everywhere, but they are filled with errors, and these errors are likely to cause faulty inferences. We should distrust performance measures, but we cannot ignore them because they are powerful motivators that can produce dramatic improvements in human and organizational performance.

Key words: measurement, methodology, performance

1. Introduction

Organizational performance has become a dominant theme in contemporary, industrialized societies. Business firms issue performance reports at least quarterly, and so newspapers and television report measures of financial performance for many firms every day. Executives' compensation depends on the numbers in these performance reports, and executives appear in the media calling attention to good performance or rationalizing poor performance. Many academic studies emphasize performance. Studies of business strategies use measures of financial performance as dependent variables; studies of work and workers use measures of job performance or job satisfaction as dependent variables. Organization theoretic studies tend to use performance measures as independent variables, as do some studies of employment turnover. However, a number of scholars have expressed doubts about the use of "performance" as a variable in management studies.

My own concerns focus on methodological issues posed by error in measures of performance and the meaning of performance data. To appreciate the importance of error in measures of performance, one should start by

★This commentary builds on a presentation at the 2004 meeting of the Academy of Management. The presentation was part of a symposium organized by Alfred Kieser and titled "Do studies of performance create actionable knowledge?"

acknowledging the biases of statistical analyses in the social sciences. A central issue is that statistically significant correlations are ludicrously easy to obtain because the actual distributions of correlations do not match the assumptions made in significance tests. Jane Webster and I assembled a database of over 13,000 correlations that were computed in studies published in *Administrative Science Quarterly*, the *Academy of Management Journal*, and the *Journal of Applied Psychology* (Webster and Starbuck, 1988). We took all the correlations among all variables observed in studies, not merely the correlations relating to hypotheses. In all three journals, the mean correlation was close to $+0.09$ and the distributions of correlations were very similar. Sixty nine percent of all correlations are positive, and 65% of all correlations are statistically significant at the 5% level. Finding statistical significance is very easy in this population of correlations. Choosing correlations utterly at random, a researcher has 2-to-1 odds of finding a significant correlation on the first try, and 24-to-1 odds of finding a significant correlation within three tries (also see Hubbard and Armstrong, 1992). Furthermore, the odds are better than 2-to-1 that an observed correlation will be positive, and positive correlations are more likely than negative ones to be statistically significant.

Thus, researchers – at least the researchers who are using methods of statistical analysis that rely on squared errors – should regard statistical significance as a treacherous indicator of the existence of theoretically meaningful relationships. Utterly random combinations of variables may appear to correlate closely. In fact, Peach and Webb (1983) demonstrated that random combinations of macroeconomic variables produce multiple correlation coefficients that appear just as large as the ones that economists report as demonstrations of the effectiveness for their macroeconomic models. Thus, errors in variables cause statistical procedures to identify incorrect associations among variables, not merely incorrect coefficients for relations.

Errors in independent variables tend to be more problematic than errors in dependent variables, so errors in measures of performance are less troublesome when performance is a dependent variable, although errors in both dependent variables and independent variables can contribute to mistaken inferences (Rousseeuw and Leroy, 1987). Figures 1 and 2 illustrate why errors in independent variables cause more trouble. Figure 1 shows some data and a line fitted to these data. One instance of the independent variable has been displaced from its correct value, possibly because of a data-entry error. As you can see, the regression line is very different from the line that would have been computed with correct data. Figure 2 shows the same original data, but this time, exactly the same error occurred in one instance of the dependent variable. Although the regression line is not the line that would have been computed with correct data, the error has distorted the regression line without making it wildly inappropriate.

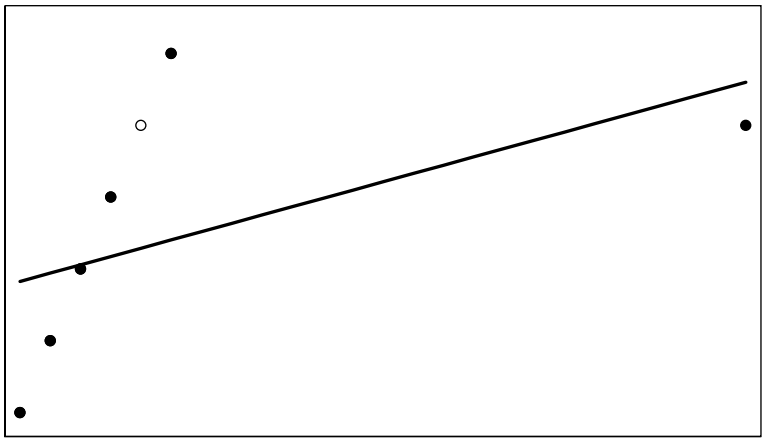

Figure 1. Error in an independent variable

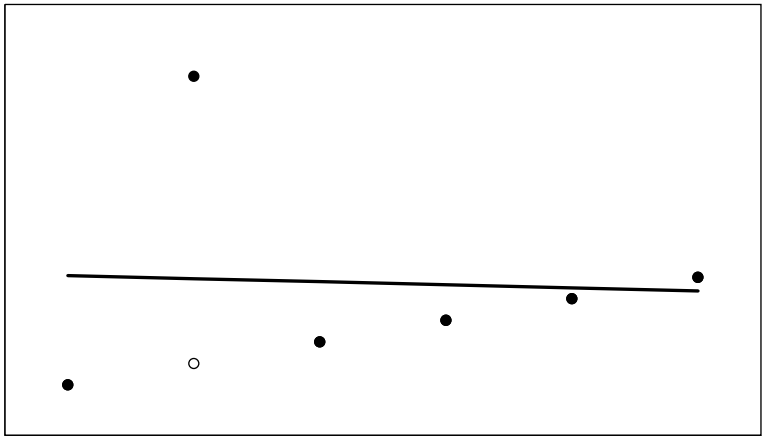

Figure 2. Error in a dependent variable

2. How Clean are the Data?

An obvious implication is that one is likely to discover spurious statistically significant relationships if even one variable incorporates large errors. Large errors in two or more variables make spurious relationships very likely. Despite the conventional assumptions of statistical models, data errors do not reduce correlations to zero. Two variables that contain large errors may (indeed, likely will) correlate significantly. For squared-error statistical inferences to yield theoretically meaningful relations, the data need to be rather clean.

However, the data used in management research may not be clean. One commonplace source of statistical data is a large database such as Compustat.

San Miguel (1977) found a 30% error rate in Compustat's reporting of R&D expenditures. These errors arose both from firms' reporting and from Compustat's processing (e.g., data-entry errors). Similarly, Rosenberg and Houglet (1974) examined the stock prices reported by Compustat and by the Center for Research in Security Prices at the University of Chicago. They (p. 1303) remarked, "There are a few large errors in both data bases, and these few errors are sufficient to change sharply the apparent nature of the data."

Error rates on the order of 20–30% pose serious problems for squared-error statistics. One criterion that statisticians use to evaluate regression methods is their "breakdown point." The breakdown point for ordinary least-squares regression is one observation. That is, a single defective observation can turn an ordinary regression into garbage. But there are alternative, robust statistical methods that can tolerate error rates approaching 50%. Thus, it may be possible for researchers to cope with the errors in large databases by adopting robust statistical methods.

A second source of data is people – data obtained either through interviews or questionnaires, including many of the data in government databases. Payne and Pugh (1976) reviewed scores of studies in which researchers had asked firms' members to characterize their firms' structures and cultures. They found that members' beliefs about their firms correlate very weakly with measurable characteristics of their firms. Likewise, John Mezias and I (2003) found similar results in two attempts to assess the accuracy of managers' perceptions. Only three-eighths of managers have perceptions that are fairly accurate, and the accuracy of managers' perceptions does not correlate with their job specializations or experience. That is, people who are supposed to know things are not more likely to perceive them accurately than people who are not supposed to know them. Further, a surprising (to me) fraction of managers have very, very erroneous perceptions; some of the perception errors go up into thousands of percent.

The very high error rates in managers' perceptions may be too large for research methods to conquer. There are, so far as I have been able to determine, no statistical techniques that will produce accurate analyses when more than half of the data are unreliable.

3. Are Errors Truly Uncorrelated?

A further problem with data obtained from people is that the errors in their data are correlated. It is a standard assumption of statistical models that each equation has errors that are uncorrelated with the variables. But this assumption is generally implausible as a description of people's perceptions because human brains impose logical order. That is, a person with an erroneous perception of one variable is likely to have an erroneous perception of other variables that the person perceives as being logically related to the first

variable. Indeed, human brains invent perceptions of events that never occurred but that seem as if they should have occurred. So someone who perceives an organization as, say, stable and orderly is very likely to also perceive the organization's environments as placid, whereas someone who perceives an organization as changing and disorderly is very likely to also perceive the organization's environments as turbulent. A study that obtains data about organizational characteristics from the same people that supply data about the organization's environment is almost certain to discover relationships that have no basis beyond a mythology constructed by common sense. Such mythological relationships are likely strongest when researchers obtain data from respondents at one time and through one method. By including items in a single questionnaire or interview, researchers suggest to respondents that they ought to see relationships among these items.

4. Do Performance Data Mean What They Appear to Mean?

Some years ago, I edited a manuscript that reported in passing that the occupations with high levels of job satisfaction include coal mining and garbage collection. This observation caught my attention because these are such extremely dangerous and unpleasant jobs. Why would people report great satisfaction from jobs that are dangerous and unpleasant? The answer, of course, lies in the nature of job-satisfaction data. These are not the reactions of people who have tried several different jobs – medical practice, stenography, farming – in addition to coal mining and garbage collection. These are the reactions of people who have remained coal miners and garbage collectors and who may never have experienced other jobs. We can surmise that they would not have remained coal miners and garbage collectors if they had the option of changing to other occupations that they preferred. They are in jobs that exact high costs of them, and they can remain in those jobs only if they evaluate their benefits from these jobs as being higher than their costs. Were we to use such findings unthinkingly, we might design jobs that are extremely dangerous and unpleasant.

I believe this illustrates a general principle about measures of performance that cut across people: There are no reliable ways to compare the satisfactions-dissatisfactions of different people (Elster and Roemer, 1993). If you and I each eat half of the same apple, how can we decide whether you enjoyed the apple more than I did? Payne and Pugh (1976) concluded that different members of an organization disagree so strongly with each other about organizational properties that it makes no sense to talk about average beliefs. In a similar vein, Friedlander and Pickle (1968) looked at the evaluations of organizational effectiveness by diverse stakeholders. They found considerable disagreement among the evaluations of owners, employees, communities, customers, suppliers, and creditors.

5. But Performance Measures Can Change Behavior

In December 2003, the *New York Times* reported that 193 American cities that have populations over 100,000 have crime rates higher than those of New York City and only five such cities have lower crime rates. Of course, it has not always been this way. Two or three decades ago, New York City had one of the highest crime rates in the U. S. So what happened?

Well, several things happened, but there is wide agreement that one of the most important things that happened was Compstat. Compstat is a management approach that emphasizes frequent, current measures of performance (Smith and Bratton, 2001). Police officers receive daily reports of the numbers of serious crimes in their precincts; they are required to submit plans for reducing these numbers; and they are held accountable if they do not reduce these numbers. It seems that performance measures can produce performance.

Thus, we should look upon performance measures with great ambivalence. There are many reasons to view performance measures with skepticism, but performance measures motivate people to perform, and erroneous performance measures can motivate efforts that waste efforts or produce unexpected results. So we must strive to produce better measures of performance – measures with fewer, smaller errors and measures that are closer to the phenomena we want to influence. The right measures of performance can produce dramatic improvements in human and organizational performance.

References

Elster, Jon and John E. Roemer: 1993, *Interpersonal Comparisons of Well-Being* (Cambridge: Cambridge University Press).

Friedlander, Frank and Hal Pickle: 1968, ''Components of Effectiveness in Small Organizations'', *Administrative Science Quarterly* 13: 289–304.

Hubbard, Raymond and J. Scott Armstrong: 1992, ''Are Null Results Becoming and Endangered Species in Marketing?'', *Marketing Letters* 3(2): 127–136.

Mezias, John M. and William H. Starbuck: 2003, ''Studying the Accuracy of Managers Perceptions: A Research Odyssey'', *British Journal of Management* 14: 3–17.

Payne, Roy L. and Derek S. Pugh: 1976, ''Organizational Structure and Climate'', in M.D. Dunnette (ed.), *Handbook of Industrial and Organizational Psychology* (Chicago: Rand McNally), pp. 1125–1173.

Peach, James T. and James L. Webb: 1983, ''Randomly Specified Macroeconomic Models: Some Implications for Model Selection'', *Journal of Economic Issues* 17: 697–720.

Rosenberg, B. and M. Houglet: 1974, ''Error Rates in CRSP and Compustat Data bases and their Implications'', *Journal of Finance* 29: 1303–1310.

Rousseeuw, Peter J. and Annick M. Leroy: 1987, *Robust Regression and Outlier Detection* (New York: Wiley).

San Miguel, Joseph G.: 1977, ''The Reliability of R&D Data in Compustat and 10-K Reports'', *Accounting Review* 52: 638–641.

Smith, Dennis C. and William J. Bratton: 2001, "Performance Management in New York City: Compstat and the Revolution in Police Management," in Dall W. Forsythe (ed.), *Quicker, Better, Cheaper? Managing Performance in American Government* (Albany, NY: Rockefeller Institute Press), pp. 453–482.

Webster, E. Jane and William H. Starbuck: 1988, "Theory building in industrial and organizational psychology," in Cary L. Cooper and Ivan T. Robertson (eds.), *International Review of Industrial and Organizational Psychology 1988* (London: Wiley), pp. 93–138.