# Merged Datasets:
## An Analytic Tool for Evidence-Based Management

**Palmer Morrel-Samuels**
**Ed Francis**
**Steve Shucard**

"Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted."—Albert Einstein

Effective evidence-based management requires analyzing data from a broad array of sources[1] and conducting carefully designed pretest-posttest comparisons.[2] However, our experience suggests that few businesses take that process to the next level by building merged datasets that can be used for rigorous pretest-posttest comparisons and meaningful statistical analyses. When data are merged from diverse independent sources across a business, researchers can then make evidence-based decisions and run pilot tests with a precision, speed, and breadth that have not been practical until now. Evidence-based management becomes especially useful when researchers build large merged datasets that are progressively linked with each other over time and that include a time series of measurements reflecting past, current, and subsequent performance. This article provides the guidance and background to aid researchers who want to build these merged datasets without further outside assistance.

Researchers in psychology,[3] medicine and health care,[4] education,[5] public health,[6] computer science,[7] business,[8] and numerous interdisciplinary fields have used and advocated aspects of evidence-based decision making for decades; often, but not always, while citing the respected traditions from which that approach has emerged: quasi-experimental analysis[9] in the behavioral sciences, and evidence-based medicine.[10] A brief word about the theory behind

The authors would like to thank our colleagues and clients for their indispensable assistance on this research.

evidence-based management, as well as its origins, will put our expansion of the work into its proper context.

## A Brief Note on Theory and Origins

The theoretical underpinnings of our approach trace an interesting story that is rarely acknowledged. Evidence-based management derives its name and method from evidence-based medicine—a field usually attributed to a loosely organized consortium of physician-educators. The consortium's initial report is frequently cited as the starting point for evidence-based medicine.[11] However, few realize that evidence-based medicine was developed to improve the education of physicians, and that the method's assumptions come from a theory of adult learning articulated by Neame and Powis.[12] Although much of the work in evidence-based management[13] is unapologetically atheoretical (as is the related work in quasi-experimental analysis, analytics,[14] and business intelligence[15]) it is occasionally helpful to recall that evidence-based management, like its precursor in the medical area, rests on adult learning theory.

## A Merged Dataset: The Essential Tool

The critical tool in evidence-based management is a large merged dataset that welds together a multitude of "hard" performance metrics and "soft" survey data measuring the corporate culture. The familiar and rudimentary uses of such a dataset are measuring performance across the organization and documenting change. The more advanced and less common uses are measuring the impact of programs,[16] identifying and quantifying linkages,[17] capitalizing on positive deviance,[18] discerning emerging trends masked by "background noise" from irrelevant factors,[19] evaluating the effect of specific leadership styles,[20] measuring the impact of communication on profit,[21] or computing the Return on Investment (ROI) of complex interventions where many variables exert their influence simultaneously.[22] The primary process consists of rigorous, methodical pretest-posttest comparisons that many readers typically associate with medical research, public health, or behavioral science.

Palmer Morrel-Samuels teaches survey design and research methods at the University of Michigan in Ann Arbor. He is also CEO of Employee Motivation and Performance Assessment and president of the Workplace Research Foundation. <palmer@umich.edu>

Ed Francis is director of E-Learning at Root Learning Inc. in Sylvania, OH.

Steve Shucard is a former manager in the IT and aeronautics industries; he is currently an affiliate at the Workplace Research Foundation.

The first step in building a merged dataset is to locate and combine ("merge" in some computer languages, "join" in others)[23] all the important databases that track a corporation's performance, resources, profits, and expenditures, regardless of their scope, location, and focus. That is, these data are moved across the enterprise into one repository, where each database is aggregated (averaged) by a common indexing variable based on one common unit of analysis (e.g., the organization's ID number) and cross-indexed

by time (viz., hour, day, week, month, or year) so that all information can be indexed to a date and a business unit within the company.

Additional rows of new data are concatenated onto the bottom of the dataset at regular intervals (e.g., every week). The dataset also grows by adding new columns containing lagged data that track performance during the previous month and the next month. Accordingly, four data manipulation procedures (merging, indexing, concatenating, and lagging) are used to build the unified dataset. Note that lagging is a two-part process wherein current data lags backward in time (enabling a comparison between this month's and the previous month's performance) as well as forward in time (enabling a comparison between this month's and the next month's performance). These four procedures are really quite straightforward, as Figure 1 illustrates.

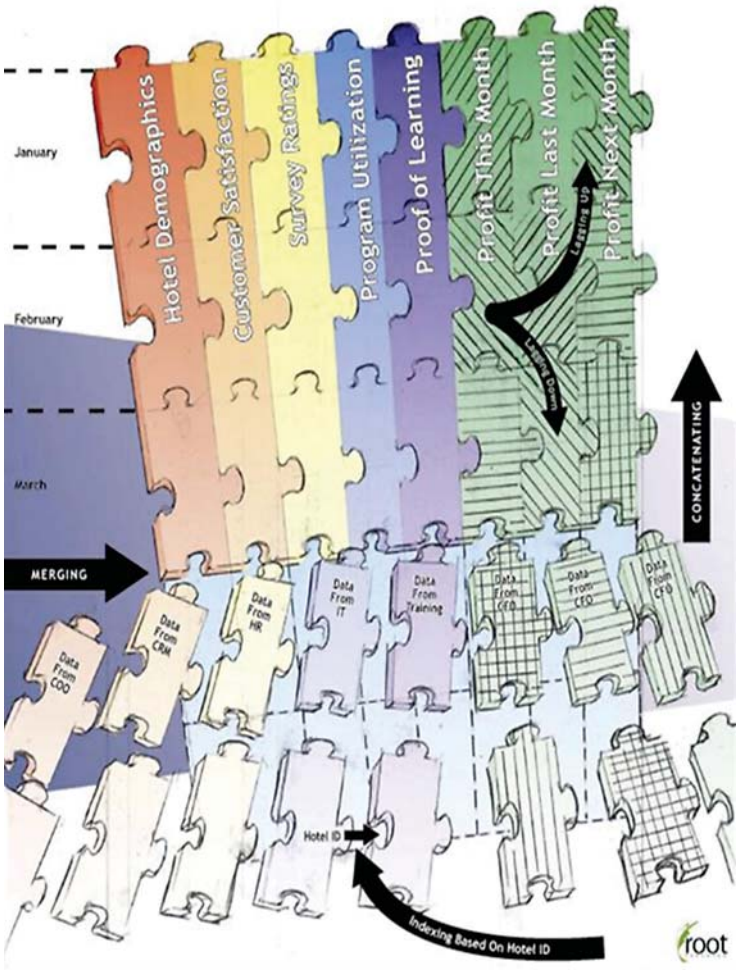
People unfamiliar with quantitative methods may believe they are already merging data in their Profit & Loss statement—which seems true enough at first glance. However, the dashboard or scorecard from a conventional P&L summary is purely descriptive, generates no unified dataset, and cannot be used to measure causal linkages. Unlike a conventional P&L summary, a typical analysis of a merged dataset provides rich diagnostic and prescriptive information that comes from every domain of the corporation. Proper use of a merged dataset makes it possible to diagnose root causes, measure impacts, evaluate the effectiveness of corporate initiatives, prescribe interventions, and forecast performance. The principles behind this kind of analysis are not new, and often, for example, hinge on accessing and applying the proper covariates for a multivariate statistical analysis.[24] However, the utility, precision, ease, and scope certainly *are* new—especially in the business world, where analytic methodology has lagged behind similar work in the behavioral sciences.

## Merged Datasets: Two Brief Examples

Merged datasets can facilitate decision making in a broad range of circumstances, even when conditions are less than ideal. For example, a merged dataset recently played the key role in resolving a contract dispute between Tower, one of the nation's largest vehicle frame manufacturers, and Lamb, a company that builds automated welders. Tower had sued Lamb for $36M, claiming that inherent defects in Lamb's welding robots were causing extended downtime. Lamb was able to defend itself by building a merged dataset that joined data from four million downtime events during a two-year period with data on absenteeism, salary, staffing levels, staff expertise, and a host of "hard" and "soft" variables drawn from Tower's own enterprise.

The analysis proved that the duration of robot downtime had very little to do with the robots themselves. Ninety-eight percent of the variance in downtime was accounted for by managerial variables such as salary, absenteeism, burnout of the workforce, and the workers' level of expertise on the assembly line—the latter being an important "soft" metric evaluated independently by subject matter experts using a standard double-blind rating procedure. Lamb's

**FIGURE 1.** The Four-Step Process for Building a Merged Dataset




welding robots were absolved, and the company was able to reach a favorable settlement out of court.

Merged datasets also have a good track record in mature organizations where current business is being challenged by an exponential growth in information. For example, until recently, Walt Disney World had used several dozen isolated datasets to track employees, hotel guests, restaurant customers, and ride patrons. When these datasets were merged into a single database, it became possible to run a rigorous analysis of staff retention, customer satisfaction, and waiting times at rides and restaurants. The analysis revealed that an entirely unanticipated variable was a predictor of staff retention: intrinsic motivation (not employee satisfaction, as presupposed). Good staff retention (not reports of "having a sense of fun" at work, as assumed), in turn, was the best predictor of

short waiting times at the restaurants and rides. Disney now has an ongoing process to merge datasets so that staff development initiatives and recruiting efforts are combined in one database that contains information from several previously isolated domains.

Both of these examples have a number of features (discussed in the following section) that are common to organizational research with merged datasets.

## A Basic Tool and a Basic Process

Our approach to evidence-based management draws heavily from standard methods in organizational research. Like all applied research, the stress is on using standardized tools and replicable processes. These common elements are described directly below.

- ***The Basic Tool:*** The main tool is the central dataset itself, which is compiled from diverse independent spreadsheets that are typically updated on different schedules and maintained by numerous champions who are autonomous, independent, and decentralized. These decentralized datasets have several features in common:
  - — They contain both quantitative and qualitative information.
  - — They track actual as well as perceived performance on tasks that are critical to the organization's financial viability.
  - — They grow on a regular schedule (e.g., monthly) so that the dataset is continuously supplemented by new data.
  - — They usually contain data that vary in precision, objectivity, and business utility.
  - — Datasets are organized by organizational entities (e.g., divisions) and by time periods.
  - — They typically, although not invariably, combine data from the four domains outlined in work on the balanced scorecard:[25] *executive* (e.g., financials, production, quality); *customer* (e.g., customer satisfaction, customer retention, market share of the customer's total expenditures [often called "share of wallet"], and complaints); *employee* (e.g., staff retention, EEO lawsuits, grievances, employee survey, performance reviews, skill assessments, and data on voluntary training courses); and *shareholder* (e.g., stockholder return, Income Available for the Common Stock or IACS, Return On Investment or ROI, and Earnings Before Interest, Taxes, Depreciation and Amortization [or EBITDA]). Despite the breadth of measures available in many corporations, some researchers draw data from just a few of the balanced scorecard domains, perhaps because a good deal can be learned by analyzing the interrelations between measures within any given domain.[26]

- ***The Basic Processes:*** Organizational research with a merged dataset entails a methodical, analytic process in which performance is compared to an objective and appropriate standard. In almost all cases, those standards come from a well-controlled pretest-posttest comparison where two or more similar groups are compared before and after a specific intervention. Analyses based on pretest-posttest comparisons are a critical part of evidence-based decision making (often associated with the term "treating the organization as a prototype" in work by Pfeffer and Sutton[27]). Pretest-posttest comparisons have the following features:
  - — They are practical, primarily because actual business performance is evaluated both before and after an intervention is made available to employees or customers.
  - — They are objective, favoring no clique within the corporation.
  - — They are unobtrusive, so the comparison process does not interfere with the organization's core business.
  - — They are replicable and can be subsequently scaled to fit larger or smaller parts of the organization, because they typically stress standardization and consistency.
  - — They are methodologically rigorous and ideally follow a conventional pretest-posttest design where employees or customers are categorized into a number of approximately equivalent groups that are then randomly assigned to a treatment or a control condition—with only the former group having access to the intervention being tested. (In cases where a standard baseline cannot be drawn from a pretest, some companies substitute a benchmark to fill this role.)

## Descriptive Outcomes vs. Diagnostic Outcomes

Many companies already have a rudimentary analytics program that uses basic data mining to build a dashboard or a balanced scoreboard of important metrics. These basic analytic tools make it possible to track changes over time and summarize the corporation's current performance, but allow nothing more than simple descriptive statistics (e.g., averages and ranges) and outcomes that are purely *descriptive*. However, in more sophisticated applications, data mining and analysis are much more rigorous, and the outcomes are *diagnostic*. In these more sophisticated research programs, the goal is to diagnose root causes, cross-validate the assessments that lead to specific job actions, measure the impact of interventions, identify subtle emerging trends, forecast performance under different scenarios, and prescribe remedial interventions to solve specific problems. To coin an analogy, descriptive analytic initiatives are to diagnostic analytic initiatives as taking a patient's pulse is to running a full battery of diagnostic tests—ultrasound, x-ray, blood analysis, and MRI—to get a full picture of the patient's current health.

## Two Specific Diagnostic Outcomes

Two major diagnostic outcomes result from using a merged dataset as part of effective evidence-based management: understanding the effect of a specific intervention and understanding a complex relationship between two variables (e.g., customer satisfaction and market share). Both outcomes are far more informative than the descriptive outcomes of a conventional dashboard or scorecard because they help us understand the causal linkages between variables.

- ***Measuring the impact of a specific intervention or program:*** These efforts usually involve a pretest-posttest comparison. In the behavioral sciences, public health, organizational psychology, and similar disciplines, multivariate inferential statistics (i.e., statistics using more than two variables, and generating *p* values, such as multiple regressions, MANOVAs, and so forth) are an essential part of such evaluations; however, in the business world, some pretest-posttest comparisons are unfortunately made without running the statistical tests that differentiate genuine differences from those caused by chance variation alone. At a minimum, these pretest-posttest comparisons require simple statistics such as correlations and/or t-tests, both of which are available in common programs such as Microsoft Excel.
- ***Identifying or measuring a complex driver:*** These efforts focus on causal linkages that are hard to measure, in part because they are deeply entangled in complex social systems, and also because their analysis usually requires considerable reliance on survey data and statistical tests. (It is important to note that many of the advanced statistical tools necessary for analyzing complex linkages, such as multiple regression, are now readily available in Microsoft Excel.) Complex drivers are common in organizational research, where *confounding variables* (i.e., unmeasured factors driving one or more critical variables in a model) can impede interpretation, where *mediating variables* (i.e., factors interposed between two important variables, just as mastery might lie between years of education and salary in a model of income), and *moderating variables* (i.e., factors that dramatically change the manner in which one variable might affect another, just as gender might change the effect of exercise on some health outcomes) and where impacts can be broadly distributed as a diffuse characteristic (such as an emphasis on personal accountability) that seems to permeate much of a corporate culture. Applied research that uses merged datasets to analyze complex drivers usually measures entities like the following: the magnitude of a known strong linkage (e.g., between employee engagement and staff retention); the impact of a weak and poorly understood linkage (e.g., between teamwork and defect rate); a weak linkage embedded in a complex system (e.g., ROI of staff retention bonuses); a linkage that is just beginning to grow in strength (e.g., from a newly expanded customer service program); or a weak impact where an outcome may risk legal complications (e.g., because it could involve discrimination). While some statistical analyses are beyond the scope of

non-specialists, the greater availability of statistical tools on desktop computers during the last decade has unquestionably made it easier to conduct organizational research of considerable scope and value.

Both of these diagnostic outcomes bring the researchers face-to-face with a vexing problem: It is often exceptionally difficult to convince non-statisticians in the business world about the magnitude and stature of a causal linkage. One sector of the audience seems inclined to mistakenly assume that every plausible impact is large and universal, while another sector seems unable to overcome its skepticism about the value of any quantitative research. The problem, we suspect, stems from the fact that too many researchers rush to offer proofs, but fail to think carefully about what is required to prove a causal linkage. In the business world especially, those specious arguments are typically weakened by overstating the claims or by pointing to misleading graphs that oversimplify, mislead, or distort.

However, with the proper forethought and attention to detail, evidence-based programs using merged datasets (and equally convincing graphs) can provide well-documented and appropriately cautious arguments that suggest causality even to an audience that has no special affection for statistics or quantitative analysis.

## Some Best-Practice Case Studies

Our case studies are organized around three critical questions: What is being measured? How is control exerted to eliminate irrelevant factors? When is the pretest-posttest comparison made? Because the possible responses to these questions are, for the present purposes, all dichotomous, our case studies fall into eight possible groups: We measure either a program or a "soft" driver that tracks an important aspect of the corporate culture, such as leadership, ethics, or communication; we provide control either by using a randomly designated control group that receives no treatment, or by using a statistical control variable that partials out (i.e., "controls for") potentially confounding factors by functioning as a covariate; and we schedule pretest-posttest comparisons on either a cascading schedule (where treatments and measurements occur continually at different times) or a classic pretest-posttest design with random assignment to a treatment group where all pretests are given simultaneously, all treatments administered simultaneously, and all posttests completed simultaneously. For brevity, our case studies will describe only the four most common of these eight combinations.

### *Case 1: Measuring Program Impact with a Randomized Controlled Trial*

At Panda Restaurant Group, we tested the impact of Root Learning's electronic learning modules—self-paced computerized tutorials—on profit, productivity, and customer satisfaction. We adopted a classic pretest-posttest design with a treatment group and a non-treatment group determined by random

assignment. Specifically, restaurants were randomly assigned to either the treatment or the non-treatment control group. We made sure each group had a similar share of different restaurant venues (e.g., mall vs. free-standing) and sub-venue (e.g., a building-end location vs. an internal location). The treatment group and the control group each contained 16 restaurants. Restaurants in the treatment group had access to the electronic learning tutorials ("e-learning modules") developed by Root, whereas restaurants in the control group did not. The analysis ran using a classic pretest-posttest design with random assignment; that is, the treatment group and the control group were determined by random assignment and both were tested twice, simultaneously—once before treatment began, and once after treatment was completed.

In an unanticipated wrinkle, some restaurants in both the treatment group and the control group received some extra attention during the study when several high-level managers and executives made a few unannounced site visits during the testing period. Nevertheless, final results suggested that these visits produced only a small and temporary improvement in restaurant performance, a finding that is consistent with a considerable body of published research.

The pretest-posttest component of the analysis was critical. It specified that the performance of both groups of restaurants be simultaneously evaluated with identical metrics both before and after the treatment period. In this case, the performance metrics consisted of a battery of metrics tracking profit, number of sales transactions, customer satisfaction, and a host of other key variables. A preliminary part of the analysis—and, in fact, the part that makes such designs informative—was an evaluation of the seasonal trends. During the month when the e-learning modules were available to the treatment group, the 1,000 restaurants in the Panda chain (on average) saw gross sales fall, all key productivity ratios decline, and customer satisfaction rise. The same seasonal pressures were also doubtlessly affecting the 32 restaurants in the treatment and control groups. However, both the treatment group and the control group saw performance improve somewhat during the test period. It is possible that the slight performance improvement was partially the result of the Hawthorne Effect, in which performance was elevated by the extra attention employees received during the management site visits.

Beyond the modest impacts from seasonal trends and site visits, the central finding of the study was clear: Performance improved more in the treatment group restaurants than it did in the control group restaurants, and the magnitude of that difference was economically and statistically significant. Specifically, the total number of sales transactions rose in both groups, but the number of transactions increased significantly more in the restaurants where employees used the e-learning modules. Furthermore, the total gross sales also rose in both groups, but gross sales increased significantly more in the restaurants where employees used the e-learning modules. Similarly, productivity ratios improved in both groups, but again, total productivity increased significantly more in the restaurants where employees used the e-learning modules. Surprisingly, while

these treatment group restaurants were doing substantially more business, making more money, and being more productive, their customer satisfaction scores also rose, and they rose to a significantly higher level than that measured in the control group restaurants.

The evidence is straightforward, consistent, and compelling that the treatment group restaurants outperformed the control group. Because random assignment was used to select the two groups, we can be quite certain that the best explanation for the sustained and pervasive improvement in the treatment group restaurants is utilization of Root's e-learning modules.

### *Case 2: Measuring Program Impact with a Non-Randomized Comparison Group*

Although randomized controls are the norm in medical quasi-experimental research, most research in organizational settings uses statistical control variables (rather than a randomized experimental manipulation), an approach that we have seen work well in hundreds of organizations. A few brief examples will suffice.

In one typical example, EDS asked Root to provide a series of train-the-trainer discussion groups that used a set of Learning Map® modules to help the 60,000 employees in EDS-GM improve customer satisfaction. As part of a simultaneous but independent contract, EDS asked EMPA (the first author's company) to provide an objective program evaluation that would determine whether Root's intervention was having the desired impact. EMPA used a standard pretest-posttest comparison design with statistical control variables (to "control for" the effect of erratic program attendance, for example), so that survey data and objective performance metrics could be compared at two times, namely, the period of the program's intervention and six months later. Specifically, the post-test survey asked each participant whether they had attended the Learning Map® discussions and whether they had applied the methods advocated in those discussions during the last six months.

These statistical control variables not only allowed us to compare employees who did participate in the discussion groups with those who were absent, but more importantly, to distinguish those who said they applied what they had learned in the discussions from those who did not.

This statistical control functioned as a critical covariate that separated the participants from employees who found a way to avoid these mandatory training sessions. The issue, of course, is that employees may have missed the training for any number of reasons. For example, they may have already felt closely attuned to customer service, or may have been skeptical about virtually all new company initiatives, or may have thought that customer satisfaction in their unit was too poor to benefit from conventional remediation unless and until sweeping changes were instituted in the product line or the service warranty.

So the critical comparison in this case was not between those who did and did not attend, but between those who attended and applied the information versus those who attended but admitted that the discussions did not change

the way they worked. A statistical analysis demonstrated the benefit of applying the method advocated in the discussions: Employees who did not attend the discussions saw no subsequent increase in customer satisfaction. Likewise, employees who attended the discussions but did not apply the approach advocated in those discussions saw no subsequent increase in their customer satisfaction scores. However, employees who attended the Learning Map® discussion groups *and* applied the methods advocated in those discussions saw their customer satisfaction scores increase by 10 percent during the six months between the pretest survey and the posttest survey. Moreover, the more those employees applied that information, the more customer satisfaction improved—a result that is rarely seen in the absence of a genuine causal relationship. By contrast, employees who simply attended without applying the information—like employees who did not attend the discussion groups at all—saw no change in customer satisfaction. The results stand as a reasonably clear example of how the quasi-experimental approach can be used to "control for" confounding variables when random assignment to a control group is not feasible (see sidebar "A Primer on Quasi-Experimental Analysis").

### *Case 3: Measuring a Driver's Impact with a Statistical Control*

The same approach described above can be used to measure the impact of complex drivers in the corporate culture, such as teamwork, customer orientation, or intrinsic motivation. For example, for four years, Fallon Clinic—the largest group practice in central Massachusetts—administered an annual survey to all the employees in its 34 clinics. Each year, a statistical analysis of the merged dataset had suggested that communication between staff was a powerful driver of patient satisfaction and profit. As outlined in a recent article that appeared in Physician Executive,[28] results of the survey showed that the CEO's efforts to change and improve the way the whole organization handled communication did indeed have their desired effect: Communication improved each year of the study.

More important, and more valuable from a financial perspective, those improvements in communication seemed to facilitate additional benefits such as enhanced patient satisfaction and increased profit. (Results for profit and patient satisfaction were similar; we focus on the link to financial performance in this brief discussion.) Specifically, in our analysis of the merged dataset for all years between 2002 and 2005, we found all four hallmarks of a causal linkage between communication and profits: communication is contemporaneously associated with profit; communication this year predicts profits next year; no other available "soft" variable that we measured—e.g., perceived teamwork, leadership, resources, pay and benefits, or fairness—had a statistically significant linkage to profit; and the greater the increase in communication between last year and this year, the greater the increase in profit between this year and the next. Archival research shows that, when presented together, these four hallmarks—association, prediction, exclusion of alternatives, and dose dependence—are apparently sufficient to convince non-statistician decision makers in court (i.e., judges and jurors) that a causal linkage does indeed exist;[29]

## A Primer on Quasi-Experimental Analysis

Quasi-experimental analysis was developed by research psychologists in the early 1960s who wanted to formulate some methods for non-laboratory research, where strict experimental control is not available.[a] Quasi-experimental analysis is now one of the main analytic methods used by clinical researchers in medicine, by quantitatively oriented organizational psychologists, by researchers using evidence-based methods to reach decisions, and by expert witnesses examining statistical evidence of discrimination in the workplace, to name but a few examples.

Quasi-experimental analysis has two major alternatives for building control groups: a non-treatment control group selected at random, which is conventionally called experimental control, or a set of statistical control variables that partial out ("control for") the potentially confounding factors that do not easily lend themselves to experimental control (such as body weight or blood pressure). The distinction between non-treatment control groups formulated by experimental manipulation and statistical controls is critical. The former is common in double-blind trials using placebos as comparisons to actual medications; the latter is common in clinical and organizational studies where some variables such as years of education or headcount are not manipulated but merely measured as a covariate that the statistician partials out during analysis.

Researchers running a quasi-experimental analysis typically measure impacts and identify root causes by conducting a pretest-posttest comparison based on one of two schedules: a classic pretest-posttest schedule with random assignment—where the treatment group and the control group are tested simultaneously, before and after an intervention—or a cyclic schedule—where a group is compared to itself and its peer groups on a regular schedule, such as monthly or annually.

A simple analogy is helpful here: Quasi-experimental analysis is to data as time telling is to a wristwatch. A wristwatch's utility stems from the fact that we share a method for telling time in minutes, seconds, and hours. Applying this method allows us to obtain a number of desired outcomes. The basic tool, the method, and the outcomes are all related but separate entities. To extend this analogy, most companies currently use analytics with such a limited understanding that it is the equivalent of using a clock only to tell the current time of day, never realizing that the same tool could be used to synchronize plans, measure durations, and determine rates of change.

Readers who are interested in more detail should see the classic texts on this topic by T.D. Cook, and D.T Campbell, or their colleague W.R. Shadish.[b]

a. T.D. Cook, and D.T. Campbell, *Quasi Experimentation: Design and Analysis Issues for Field Settings* (Chicago, IL: Rand McNally, 1979).
b. Cook and Campbell, op. cit.; W.R. Shadish, T.D. Cook, and D.T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Boston, MA: Houghton Mifflin, 2002).

accordingly, given the fact that legal standards are presumably more stringent than most informal criteria, we can be reasonably confident that such evidence suggesting causality will be reasonably convincing in the workplace, even when absolute certainty about a causal linkage remains elusive—just as it does when interpreting the results of virtually any research.

It is interesting to note that communication is an unexpectedly important driver in many of our other analyses as well. For example, the Arthur W. Page Society (an organization of the top communications executives in the country) and Gagen MacDonald (a communications consulting firm in Chicago) commissioned a study to evaluate the predictive impact of communication on subsequent stock price. A classic pretest-posttest comparison among the 20 participating companies found that the better the communication in the corporation, the better the subsequent financial return to stockholders. These findings echo results from the National Benchmark Study,[30] which measures the relationship between corporate culture and subsequent financial outcomes at the national level.

### *Case 4: Measuring Program Impact Comparing Early and Later Adopters*

We recently completed a large impact analysis at Holiday Inn Express (HiEx), where we tested the impact of Root Learning's e-learning modules on two critical performance metrics in the hospitality industry: Revenue Per Available Room (RevPAR) and number of customer complaints. The circumstances of the analysis were dauntingly adverse. To begin with, the parent company, InterContinental Hotels Group, has franchise agreements that give the general manager who owns each HiEx facility a substantial degree of autonomy and privacy. Consequently, headquarters cannot capture many of the day-to-day metrics (such as staff headcounts) that are usually important in corporate quasi-experimental analysis.

More problematic, the e-learning modules were being rolled out on a cascading schedule. Different hotels would have different amounts of access at different times, and employees could use as many or as few of the e-learning modules as they (and presumably each hotel's general manager) chose. With HiEx adding roughly three properties to its roster of 1,600 hotels every day, the organization seemed far too fluid for any detailed analysis. However, the data-tracking system at HiEx is a model of efficiency and accuracy, so we were able to compensate for the missing and changing data by using important covariates such as number of rooms, brand, and type of location. By the end of the six-month study, we were able to obtain a clean set of performance metrics for every HiEx hotel in the system worldwide.

We merged the performance data from the hotels with data from Root Learning on the 30 e-learning modules they provided for HiEx. The data included five identical pretest and posttest questions on the content of each module. Our simple multiple-choice test of learning made it possible to measure, in a rudimentary but straightforward way, an employee's extent of knowledge both before and after taking each e-learning module. A preliminary statistical analysis found that these learning scores were normally distributed among employees and that the hotels' average scores were significantly different from each other—important prerequisites for extended statistical analysis because

they suggest that scores from employees might be high in validity and business utility.

We used merged datasets to measure the impact of e-learning modules as they were rolled out through the system. We found compelling evidence that using the e-learning modules increased revenue and lowered customer complaints. Specifically, learning scores were significantly correlated with concurrent Revenue Per Available Room, and learning scores predicted Revenue Per Available Room next month, and the greater the improvement in learning scores this month, the greater the improvement in Revenue Per Available Room next month.

Moreover, results were virtually identical when we analyzed the relationship between learning scores and customer complaints. Learning scores were associated with fewer complaints in the same month, significantly predicted fewer complaints in the coming month, and showed clear evidence of predictive dose dependence: The greater the increase in learning scores this month, the greater the drop in complaints next month.

A key aspect of the evidence-based management program at HiEx was a cascading schedule working in conjunction with statistical control. That is, each hotel used only as many new e-learning modules as employees freely chose during each day of the research period. Accordingly, our research design provided a statistical control variable for every hotel in the company, thereby making it possible to examine the precise relationship between learning scores at each hotel during each month, and objectively measured performance at the same hotel. Moreover, because data were lagged both forward and backward in time, it was possible to generate month-to-month change scores, and to analyze each month's performance while statistically controlling for performance during the previous month. The merged database and the quasi-experimental design made it possible to run thousands of rigorous pretest-posttest comparisons with one simple multivariate statistical test that examined change scores in revenue as a function of change scores in learning (while controlling for potential confounding effects associated with hotel size, season, hotel location, and a host of other factors).

Admittedly, the results of the study were weakened by the fact that we—like many researchers, for reasons that Florida and Davison summarize[31]—lacked the administrative control required to compel randomly selected hotel owners to use, or not use, the e-learning modules, a feature that would have provided numerous statistical advantages[32] and controlled for additional confounds such as the desire to learn or the number of distractions in the workplace. Nevertheless, it was still the case that the more employees learned, the greater the subsequent increase in Revenue Per Available Room—compelling evidence of some causal connection, even with an appropriate caveat. On balance, the case study provides a reasonably good example of what can be accomplished even when there is very little experimental control.

## Why Evidence-Based Management Sometimes Fails

Because companies often institute an evidence-based program hastily and with less forethought than the task requires, it is sensible to provide some cautionary examples of initiatives that failed and to outline the six major impediments that can preclude the success of an evidence-based management program.

Sometimes everything about an evidence-based initiative is perfect except the results. For example, one of our nation's largest automobile makers (a company now struggling to recover from bankruptcy) conducted a rigorous skill assessment of its 35,000 domestic engineers several years ago and was taken aback by the results. Evidence from the skill assessment clearly revealed that specific subgroups of engineers were consistently overestimating their own skill level, underutilizing voluntary training programs, and designing product lines that generated unusually high defect rates, low customer satisfaction scores, and sluggish sales. When the full results of the statistical analysis became clear, the project manager told his team, with some embarrassment, that resistance to change within the corporation was going to make it easier to implement cosmetic changes to the engineer training curriculum and "declare victory" than to address the underlying problem by compelling additional education, reassigning staff, or altering job responsibilities.

In a second case, a sweeping but poorly thought-out analytics program was implemented at one of the nation's largest, most profitable banks. The initiative was intended to compile sound evidence to support all phases of talent management in the 250,000-employee workforce, but promotions and reassignments were so numerous, and the data so voluminous, that analysts lost track of important distinctions such as the difference between conventional managerial assignments and turn-around assignments where the goal was to insert an experienced manager into a failing business unit. Accordingly, managers in turn-around assignments were unfairly penalized because their financial performance and results from their employee surveys both looked distressingly poor. In the end, the analytics initiative lost its business utility and collapsed because it failed to present data points in their appropriate and full context.

In both of these examples, evidence-based management was derailed by a poor understanding of how to interpret evidence of causality in organizational research (and arguably by a lack of courage within the project team)—problems that were exacerbated by deep silos, lack of support from the executive suite, resistance to change, territorial infighting, and, above all, a pervasive misunderstanding of the methods and goals that characterize evidence-based management.

## Seven Potential Impediments to Success

No special statistical tool is necessary for compiling or analyzing a merged dataset, so lack of sophisticated statistical analysis is not what leads to the failure of evidence-based management initiatives. Rather, we contend that resistance

to change and a limited understanding are the most potent impediments to successful implementation. In our experience, however, a good evidence-based management program can, in itself, begin to overcome institutional resistance to change and lack of understanding both inside and outside of the project team—at least to some extent. However, entirely successful implementation also requires avoidance of the following impediments:

- ***Assessments that lack validity and utility:*** If a key assessment (e.g., a customer satisfaction survey) is not valid, then the attitudes, perceptions, problems, and desires of respondents cannot be measured accurately. Accordingly, the interventions designed to help those respondents will be based on misleading data. (For more information on the importance of validity and business utility in assessment data, see "Getting the Truth into Workplace Surveys."[33])
- ***Failure to focus on employee motivation:*** If employee motivation (and its near-synonyms "employee engagement," "commitment," or "loyalty") is not a central focus of the company, then employees will be, as one client recently described it, "retired in place," so no initiatives or improvements in the employee domain—even initiatives and improvements based on sound evidence—will have a discernible impact on company performance. This issue is important because excessive cynicism in the workforce is inimical to evidence-based management: Employees need to want to improve; without that essential motivation, any new initiatives coming from even the best evidence-based program will founder on the rocky shoals of cynical indifference.
- ***Poor cooperation and communication:*** If communication in the organization is hampered by unacknowledged gaps, blockages, and chronic infighting, then poor coordination between departments and individuals will almost invariably hurt performance and preclude lasting improvements. Our experience has consistently confirmed the importance of good organizational communication[34] in evidence-based management initiatives.
- ***Lack of executive support:*** If executive support is meager or absent, progressive forces within the company will lack the resources and administrative leverage necessary to overcome resistance from the dinosaurs, snakes, and slackers who will, intentionally or not, sabotage any project that exposes poor performance.
- ***Averages that are computed inaccurately and/or inconsistently:*** Aggregating data into averages can cause an unexpectedly treacherous problem for evidence-based management programs, because computation methods differ in large organizations, and substantial misalignments can result. Averages should be "grand means" computed from raw data rather than averages of averages, a problem that can become distressingly acute when some cells in a dataset are blank. The key is to use consistent computational processes and to apply labels that explicitly differentiate grand means from averages of averages. While this is not a fatal flaw in many

organizations, we have seen its corrosive effect on evidence-based decision making, in part because persistent and vocal (albeit uninformed) criticism about minor inaccuracies can undermine even a successful program's credibility.

- ***A presentation of results that is ambiguous, confusing, or distorted:*** Minor and innocuous decisions about formatting and displaying data can have a profound impact on comprehension, perceived importance, and memorability. There are, after all, more than 70 years of published empirical research on the comprehension and recall of quantitative graphs—issues that acquire critical importance in the workplace because executive decisions are often based heavily, and sometimes even solely, on graphs.
- ***Lack of understanding about causal evidence in the applied arena:*** As work on quasi-experimentation shows, there are many good alternatives to a strictly controlled laboratory experiment. However, the guidelines for interpreting evidence of a causal linkage from such applied research necessarily must include appropriate caveats so that the strength of the evidence is neither overstated nor dismissed out of hand. And, above all, readers need to be informed that—despite the fact that perfect experimental control may not be present—a good deal about the relationships between variables can be learned.

## Attributes of Effective Evidence-Based Management Programs

We have tried in this brief analysis to encourage a clear understanding of the tools, methods, pitfalls, benefits, and typical outcomes of evidence-based management programs using a merged dataset. From our perspective, such programs are uniquely valuable. They are one of the few business initiatives that consistently get high marks for being:

- *Inclusive:* Merged datasets come from across the entire corporation—even sectors that might rarely receive prominent attention.
- *Relevant:* Merged datasets necessarily reflect the performance of a multitude of executives, managers, and individual contributors.
- *Timely:* Merged datasets stay up to date because they grow over time.
- *Useful:* Merged datasets provide objective and practical program evaluations.
- *Profitable:* Merged datasets make it very difficult for programs with unsuspected negative impacts to hide behind empty rhetoric.
- *Informative:* Merged datasets provide objective data that are indispensable for ancillary tasks such as prioritizing corporate initiatives or computing the ROI of intangibles.
- *Transparent and Easy to Understand:* Merged datasets draw upon straightforward common sense and the research on inferences that non-statisticians make when they seek to understand quantitative data.

- *Ethically Balanced:* Merged datasets provide accurate and rigorous analysis of linkages, so that profit and the corporate culture can be improved simultaneously.
- *Methodologically Sound:* Using merged datasets to facilitate evidence-based management increases reliance on best practices in applied research, such as randomized pretest-posttest comparisons, incorporation of appropriate covariates, and broad system-based thinking.

## Closing Thought

Notwithstanding Lord Kelvin's aphorism about the value of measurement ("If you can measure that of which you speak, and can express it as a number, you know something of your subject, but if you cannot measure it, your knowledge is meager and unsatisfactory."), it would be foolhardy simply to measure everything that might be important in a large corporation and hope that some insights will eventually emerge. Most of us have had the unpleasant experience of working alongside technocrats who measure everything but understand little. Instead, we advocate the judicious use of merged datasets and an ongoing commitment to evidence-based management so that sound organizational research can engender simultaneous improvement in both profitability *and* the corporate culture. We owe all involved—our employees, shareholders, customers, and society at large—nothing less.

### Notes

1. J. Pfeffer and R.I. Sutton, "Management Half-Truths and Nonsense: How to Practice Evidence-Based Management," *California Management Review*, 48/3 (Spring 2006): 77-100.
2. J. Pfeffer and R.I. Sutton, "Evidence Based-Management," *Harvard Business Review*, 84/1 (January 2006): 62-74.
3. C.F. Michaels, R. Arzamarski, R.W. Isenhower, and D.M. Jacobs, "Direct Learning in Dynamic Touch," *Journal of Experimental Psychology, Human Perception and Performance*, 34/4 (2008): 944-957.
4. S.C. Wangberg, "An Internet-Based Diabetes Self-Care Intervention Tailored to Self-Efficacy," *Health Education Research*, 23/1 (2008): 170-179; L. Bergthold, S. Olson Koebler, and S. Singer, "In Loco Parentis? The Purchaser Role in Managed Care," *California Management Review*, 43/1 (Fall 2000): 34-52.
5. P.D. Mautone and R.E. Mayer, "Cognitive Aids for Guiding Graph Comprehension," *Journal of Educational Psychology*, 99/3 (2007): 640-652.
6. Y. Yousey, J. Leake, M. Wdowik, and J.K. Janken, "Education in a Homeless Shelter to Improve the Nutrition of Young Children," *Public Health Nursing*, 24/3 (May/June 2007): 249-255.
7. R. Henderson, F. Deane, K. Barrelle, and D. Mahar, "Computer Anxiety: Correlates, Norms and Problem Definition in Health Care and Banking Employees using the Computer Attitude Scale," *Interacting with Computers*, 7/2 (June 1995): 181-193.
8. J.W. Medcof, "The Job Characteristics of Computing and Non-Computing Work Activities," *Journal of Occupational and Organizational Psychology*, 69/2 (1996): 199-212; V.A. Zeithaml, R.T. Rust, and K.N. Lemon, "The Customer Pyramid: Creating and Serving Profitable Customers," *California Management Review*, 43/4 (Summer 2001): 118-146; T.H. Davenport, J.G. Harris, D.W. De Long, and A.L. Jacobson, "Data to Knowledge to Results: Building an Analytic Capability," *California Management Review*, 43/2 (Winter 2001): 117-138.
9. J.A. Schellenberg, "The Effect of Pretesting upon the Risky Shift," *Journal of Psychology: Interdisciplinary and Applied*, 88/2 (1974): 197-200.

10. W.R. Shadish, T.D. Cook, and D.T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Boston, MA: Houghton Mifflin, 2002).
11. Evidence-Based Medicine Working Group, "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine," *Journal of the American Medical Association*, 268/17 (1992): 2420-2425; M.L. Green and P.J. Ellis, "Impact of an Evidence-Based Medicine Curriculum Based on Adult Learning Theory," *Journal of General Internal Medicine*, 12/12 (December 1997): 742-750; S.E. Straus and F.A. McAlister, "Evidence-Based Medicine: A Commentary on Common Criticisms," *Canadian Medical Association Journal/Journal De l'Association Medicale Canadienne*, 163/7 (2000): 837-841.
12. Green and Ellis, op. cit.; R.L. Neame and D.A. Powis, "Toward Independent Learning: Curricular Design for Assisting Students to Learn How to Learn," *Journal of Medical Education*, 56/11 (November 1981): 886-893.
13. R.B. Briner, "Improving Stress Assessment: Toward an Evidence-Based Approach to Organizational Stress Interventions," *Journal of Psychosomatic Research*, 43/1 (1997): 61-71; J. Pfeffer and R.I. Sutton, "Suppose We Took Evidence-Based Management Seriously: Implications for Reading and Writing Management," *Academy of Management Learning & Education*, 6/1 (March 2007): 153-155; J. Pfeffer and R.I. Sutton, *Hard Facts, Dangerous Half-Truths and Total Nonsense: Profiting from Evidence-Based Management* (Cambridge, MA: Harvard Business School, 2006).
14. M.D. Flood, "Embracing Change: Financial Informatics and Risk Analytics," *Quantitative Finance*, 9/3 (2009): 243-256; T.H. Davenport, "Competing on Analytics," *Harvard Business Review*, 84/1 (January 2006): 99-107; Davenport et al., op. cit.
15. W. Yeoh, A. Koronios, and J. Gao, "Managing the Implementation of Business Intelligence Systems: A Critical Success Factors Framework," *International Journal of Enterprise Information Systems*, 4/3 (2008): 79-95.
16. E.M. Kyrouz and K. Humphreys, "Do Health Care Workplaces Affect Treatment Environments?" *Journal of Community & Applied Social Psychology*, 7/2 (April 1997): 105-118.
17. D.E. Bowen and C. Ostroff, "Understanding HRM-Firm Performance Linkages: The Role of the 'Strength' of the HRM System," *Academy of Management Review*, 29/2 (April 2004): 203-221.
18. R. Waldersee and F. Luthans, "The Impact of Positive and Corrective Feedback on Customer Service Performance," *Journal of Organizational Behavior*, 15/1 (January 1994): 83-95.
19. R.P. Vecchio, "The Impact of Referral Sources on Employee Attitudes: Evidence from a National Sample," *Journal of Management Studies*, 21/5 (1995): 953-965.
20. J. Barling, T. Weber, and E.K. Kelloway, "Effects of Transformational Leadership Training on Attitudinal and Financial Outcomes: A Field Experiment," *Journal of Applied Psychology*, 81/6 (1996): 827-32.
21. B.S. Maini and P. Morrel-Samuels, "Cascading Improvements in Communication: Adopting a New Approach to Organizational Communication," *Physician Executive*, 32/5 (September/October 2006): 38-43.
22. R.W. Eichinger and M.M. Lombardo, "The ROI on People—The 7 Vectors of Research," unpublished manuscript disseminated by Lominger Limited LTD, Minneapolis, 2004.
23. SAS Institute, *JMP Statistics and Graphics Guide* (Cary, NC: SAS Institute, Inc., 2007).
24. J. Pfeffer and A. Davis-Blake, "Unions and Job Satisfaction: An Alternative View," *Work and Occupations*, 17/3 (1990): 259-283.
25. R.S. Kaplan and D.P. Norton, "The Balanced Scorecard; Measures that Drive Performance," *Harvard Business Review*, 70/1 (January/February 1992): 71-80.
26. We thank one of our reviewers for suggesting this point; it stands as a useful reminder that much can be learned about an organization even in the absence of perfect adherence to a comprehensive theoretical framework.
27. J. Pfeffer and R.I. Sutton, "Treat Your Organization as a Prototype: The Essence of Evidence-Based Management," *Design Management Review*, 17/3 (Summer 2006): 10-14.
28. Maini and Morrel-Samuels, op. cit.
29. P. Morrel-Samuels and P.J. Jacobson, "Using Statistical Evidence to Prove Causality to Non-Statisticians," paper presented at the Meeting of the American Psychology-Law Society, Jacksonville, FL, 2008.
30. P. Morrel-Samuels, "The National Benchmark Study: Employee Motivation Affects Subsequent Stock Price," paper presented at the Meeting of the American Psychological Association, Toronto, Canada, August 2009.

31. R. Florida and D. Davison, "Gaining from Green Management: Environmental Management Systems Inside and Outside the Factory," *California Management Review*, 43/3 (Spring 2001): 64-86.
32. S.B. Morris, "Estimating Effect Sizes from Pretest-Posttest-Control Group Designs," *Organizational Research Methods*, 11/2 (2007): 364-386.
33. P. Morrel-Samuels, "Getting the Truth into Workplace Surveys," *Harvard Business Review*, 80/2 (February 2002): 111-118.
34. P. Morrel-Samuels and B. Maini, "Cascading Improvements in Communication Throughout the Workplace," paper presented at the Annual Conference of the American Psychological Association, New Orleans, LA, 2006.