



THE ARGUMENT



THE ARGUMENT IN A NUTSHELL

There is a cultural pattern that has become ubiquitous in recent decades, engulfing an ever-widening range of institutions. Depending on taste, one could call it a cultural “meme,” an “épistème,” a “discourse,” a “paradigm,” a “self-reinforcing rhetorical system,”¹ or simply a fashion. It comes with its own vocabulary and master terms. It affects the way in which people *talk* about the world, and thus how they *think* about the world and how they *act* in it.² For convenience, let’s call it metric fixation.

A key premise of metric fixation concerns the relationship between measurement and improvement. There is a dictum (wrongly) attributed to the great nineteenth-century physicist Lord Kelvin: “If you cannot measure it, you cannot improve it.” In 1986 the American management guru, Tom Peters, embraced the motto, “What gets measured gets done,” which became a cornerstone belief of metrics.³ In time, some drew the conclusion that “anything that can be measured can be improved.”⁴

When proponents of metrics advocate “accountability,” they tacitly combine two meanings of the word. On the one hand, to be accountable means to be responsible. But it can also mean “capable of being counted.” Advocates of “accountability” typically assume that only by counting can institutions be truly responsible. Performance is therefore equated with what can be reduced to standardized measurements.

When proponents of metrics demand “transparency” they often insinuate that probity requires making explicit and visible as much information as possible. The result is the demand for ever more documentation, ever more mission statements, ever more “goal-setting.”⁵

The key components of metric fixation are

- the belief that it is possible and desirable to replace judgment, acquired by personal experience and talent, with numerical indicators of comparative performance based upon standardized data (metrics);
- the belief that making such metrics public (transparent) assures that institutions are actually carrying out their purposes (accountability);
- the belief that the best way to motivate people within these organizations is by attaching rewards and penalties to their measured performance, rewards that are either monetary (pay-for-performance) or reputational (rankings).

Metric fixation is the persistence of these beliefs despite their unintended negative consequences when they are put into practice.⁶ It occurs because not everything that is important is measureable, and much that is measurable is unimportant. (Or, in the words of a familiar dictum, “Not everything that can be counted counts, and not everything that counts can be counted.”⁷) Most organizations have multiple purposes, and that which is measured and rewarded tends to become the focus of attention, at the expense of other essential goals. Similarly, many jobs have multiple facets, and measuring only a few aspects creates incentives to neglect the rest.⁸ When organizations committed to metrics wake up to this fact, they typically add more performance measures—which creates a

cascade of data, data that becomes ever less useful, while gathering it sucks up more and more time and resources.

In the process, the nature of work is transformed in ways that are often pernicious. Professionals tend to resent the impositions of goals that may conflict with their vocational ethos and judgment, and thus morale is lowered. Almost inevitably, many people become adept at manipulating performance indicators through a variety of methods, many of which are ultimately dysfunctional for their organizations. They fudge the data or deal only with cases that will improve performance indicators. They fail to report negative instances. In extreme cases, they fabricate the evidence.

A frequent feature of metric fixation is paying for performance, that is, offering individuals or organizations financial incentives to meet quantifiable criteria. That may work in organizations that exist for the single purpose of making a profit, though as we'll see, even in these cases it is rarely effective. It works even less well in organizations in which employees are oriented to a more idealistic mission, such as schools, universities, medical practices, and hospitals. Whenever reward is tied to measured performance, metric fixation invites gaming.

Because the theory of motivation behind pay for measured performance is stunted, results are often at odds with expectations. The typical pattern of dysfunction was formulated in 1975 by two social scientists operating on opposite sides of the Atlantic, in what appears to have been a case of independent discovery. What has come to be called "Campbell's Law," named for the American social psychologist Donald T. Campbell, holds that "[t]he more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."⁹

In a variation named for the British economist who formulated it, we have Goodhart's Law, which states, "Any measure used for control is unreliable."¹⁰ To put it another way, anything that can be measured and rewarded will be gamed. We will see many variations on this theme.

Trying to force people to conform their work to preestablished numerical goals tends to stifle innovation and creativity—valuable qualities in most settings. And it almost inevitably leads to a valuation of short-term goals over long-term purposes.

In situations where there are no real feasible solutions to a problem, the gathering and publication of performance data serves as a form of virtue signaling. There is no real progress to show, but the effort demonstrated in gathering and publicizing the data satisfies a sense of moral earnestness. In lieu of real progress, the progress of measurement becomes a simulacrum of success. We'll see that in the case of the educational "achievement gap."

Because belief in its efficacy seems to outlast evidence that it frequently doesn't work, metric fixation has elements of a cult. Studies that demonstrate its lack of effectiveness are either ignored, or met with the assertion that what is needed is more data and better measurement. Metric fixation, which aspires to imitate science, too often resembles faith.

All of that is not intended to claim that measurement is useless or intrinsically pernicious. One of the purposes of this book is to specify when performance metrics are genuinely useful—how to use metrics without the characteristic dysfunctions of metric fixation.

The next chapter, "Recurring Flaws," provides a taxonomy of the most frequent types of flaws in the use of performance

metrics. Defining and labeling them will make it easier to refer back to them later. Then, in part II, we examine the origins of metric fixation and account for its spread and tenacity in spite of its frequent failures, in addition to exploring some of the deeper philosophical sources of its shortcomings. Part III comprises case studies that examine the more recent record of metrics, its successes and its shortcomings in a variety of fields, including K-12 education, higher education, medicine, policing, the military, business, and philanthropy and foreign aid. These case studies are intended to be suggestive rather than definitive. That is, they don't deal with every way in which the metric fixation manifests itself in each domain. Rather they provide concrete examples of recurring flaws and unintended consequences, as well as examples of the successful use of metrics from which we may derive lessons that can be applied in other domains. This section is followed by a brief excursus on the theme of transparency as the enemy of performance in certain realms. Finally, Part IV draws upon the preceding analysis to enumerate the unintended negative consequences of metric fixation and offer some guidelines about when and how to make use of metrics without succumbing to metric fixation.

2

RECURRING FLAWS

The drive to institute metrics often arises from the best of intentions, as a purported solution to real problems. And in some cases, as we'll see, it really *does* fulfill its promise to provide such solutions, or at least contributes to solving problems. But after decades of experience with the negative effects of metrics, as metric dysfunction threatens to cascade into yet more institutions, we should be able to anticipate the recurrent flaws. Here's a list to help identify and remember them. Of course, while we may distinguish them for purposes of analysis, these flaws often overlap in the real world.

Let's begin with problems of the *distortion* of information.

Measuring the most easily measurable. There is a natural human tendency to try to simplify problems by focusing on the most easily measurable elements.¹ But what is most easily measured is rarely what is most important, indeed sometimes not important at all. That is the first source of metric dysfunction.

Closely related is **measuring the simple when the desired outcome is complex.** Most jobs have multiple responsibilities and most organizations have multiple goals. Focusing measurement on just one responsibility or goal often leads to deceptive results.

Measuring inputs rather than outcomes. It is often easier to measure the amount spent or the resources injected into

a project than the *results* of the efforts. So organizations measure what they've spent, rather than what they produce, or they measure process rather than product.

Degrading information quality through standardization.

Quantification is seductive, because it organizes and simplifies knowledge. It offers numerical information that allows for easy comparison among people and institutions.² But that simplification may lead to distortion, since making things comparable often means that they are stripped of their context, history, and meaning.³ The result is that the information appears more certain and authoritative than is actually the case: the caveats, the ambiguities, and uncertainties are peeled away, and nothing does more to create the appearance of certain knowledge than expressing it in numerical form.⁴

Campbell's Law and Goodhart's Law are warnings about the inevitable attempts to game the metric when much is at stake. Gaming the metrics takes a variety of forms.

Gaming through creaming. This takes place when practitioners find simpler targets or prefer clients with less challenging circumstances, making it easier to reach the metric goal, but excluding cases where success is more difficult to achieve.

Improving numbers by lowering standards. One way of improving metric scores is by lowering the criteria for scoring. Thus, for example, graduation rates of high schools and colleges can be increased by lowering the standards for passing. Or airlines improve their on-time performance by increasing the scheduled flying time of their flights.

Improving numbers through omission or distortion of data. This strategy involves leaving out inconvenient in-

stances, or classifying cases in a way that makes them disappear from the metrics. Police forces can “reduce” crime rates by booking felonies as misdemeanors, or by deciding not to book reported crimes at all.

Cheating. One step beyond gaming the metrics is cheating—a phenomenon whose frequency tends to increase directly with the stakes of the metric in question. As we’ll see, as the No Child Left Behind Act raised the stakes for schools of the test scores of their pupils, teachers and principals in many cities responded by altering students’ answers on the test.

12

BUSINESS AND FINANCE

WHEN PAYING FOR PERFORMANCE WORKS, AND WHEN IT DOESN'T

“But surely,” you might think, “there is a place where pay for measured performance is appropriate, and that is in the realm of business.” Businesses, after all, exist to make money, and people work in them to make money for themselves. It makes sense, it seems, for business managers to try to elicit their employees’ greatest effort by tying their remuneration as closely as possible to their measurable contribution to making profits for the firm.

There are indeed circumstances when pay for measured performance fulfills that promise: when the work to be done is repetitive, uncreative, and involves the production or sale of standardized commodities or services; when there is little possibility of exercising choice over what one does; when there is little intrinsic satisfaction in it; when performance is based almost exclusively on individual effort, rather than that of a team; and when aiding, encouraging, and mentoring others is not an important part of the job. For sales forces,¹ or for routinized, individualized, highly focused jobs involving standardized outputs and without broader responsibilities, rewarding measured performance may well pay off. In short, as one sociologist has put it, “Extrinsic rewards become an important determinant of job satisfaction only among workers for whom intrinsic rewards are relatively unavailable.”² These

are the sort of tasks for which Taylorism (see chapter 3) was designed. There are many such jobs in any society, including a modern, technologically advanced one. But in our time, as the technologies of robotics and artificial intelligence advance, such jobs are becoming fewer and far between.³

But the salient fact is that most private-sector jobs do *not* match these criteria. And to the extent that they do not, direct payment for measured performance will be inappropriate and perhaps counterproductive.

People *do* want to be rewarded for their performance, both in terms of recognition and remuneration. But there is a difference between promotions (and raises) based on a range of qualities, and direct remuneration based on measured quantities of output. For most workers, contributions to their company include many activities that are intangible but no less real: coming up with new ideas and better ways to do things, exchanging ideas and resources with colleagues, engaging in teamwork, mentoring subordinates, relating to suppliers or customers, and more. It's appropriate to reward such activities through promotions and bonuses—even if it is more difficult to document and requires a greater degree of judgment by those who decide on the rewards. Nor is the problem assigning numbers to performance. There is nothing wrong with rating people on a scale. The problems arise when the scale is too one-dimensional, measuring only a few outputs that are most easily measured because they can be standardized.

Indeed, the academic evidence on pay for the measured performance of CEOs and other personnel is sufficiently troubling that some scholars of organizational behavior have suggested that it should simply be eliminated. And some companies are acting accordingly. Dan Cable and Freek Vermeulen of the London Business School recall many of the problems

we have explored: the depressive effect of performance pay on creativity; the propensity to cook the books; the inevitable imperfections of the measurement instruments; the difficulty of defining long-term performance; and the tendency for extrinsic motivation to crowd out intrinsic motivation. They've concluded that it might be more advantageous to abolish pay-for-performance for top managers, and replace it with a higher fixed salary. They even suggest, rather heretically, that you might not want people motivated primarily by extrinsic motivation at the head of your company: yet the more compensation is variable and linked to measured performance, the more likely that that will be precisely the sort of people you will get.⁴ And at least one of Britain's best-known investors, Neil Woodford of Woodford Investment Management, a company with £14.3 billion under management, has eliminated bonuses for the company's executives in favor of higher fixed pay, arguing that there is little correlation between bonus and performance.⁵

Forced ranking, in which managers are instructed to evaluate their employees compared to fellow employees, is another manifestation of metric fixation. It seems "hard" and "objective," but often turns out to be counterproductive. A 2006 survey of more than two hundred human resource professionals from large companies found that "even though over half of the companies used forced ranking, the respondents reported that this approach resulted in lower productivity, inequity, skepticism, decreased employee engagement, reduced collaboration, damage to morale, and mistrust in leadership."⁶

Increasing numbers of technology companies, conscious of the demotivating effect of performance rankings on the majority of their staff, are moving away from performance bo-

nuses. They are replacing them with higher base salaries combined with shares or share options, to give employees a tangible interest in the long-term flourishing of the company (while paying special rewards to particularly high performers).⁷

Yet other companies are dropping annual ratings in favor of “crowdsourced” continuous performance data, by which supervisors, colleagues, and internal customers provide ongoing online feedback about employee performance. That may be substituting the frying pan for the fire, as employees constantly game for compliments, while resenting the omnipresent surveillance of their activities⁸—a dystopian possibility captured in Dave Eggers 2014 novel *The Circle*. Yet as improvements in information technology make it easier to monitor one or another index of worker performance, it will become ever more tempting to link pay to performance, whether in the form of piece rates, bonuses, or commissions⁹—in spite of evidence of the hazards of measuring too narrowly, and of discouraging teamwork and innovation.

A great deal of corporate dysfunction comes from pay-for-performance schemes that are narrowly tailored to measure a single outcome. Problems occur both at the top and at the bottom of the corporate ladder.

For a dramatic instance among top executives, take the case of the pharmaceutical manufacturer Mylan. Though not among the largest of American pharmaceutical firms (11th by revenue and 16th by market capitalization), it had the second-highest level of executive compensation: over the course of five years ending in December 2015, its top three managers were paid over \$70,000,000 each. Over that period, its stock price rose 155 percent. In 2014 its board of directors developed

a compensation scheme for the company's top executives according to which they would be handsomely rewarded if the company's profits grew by 16 percent each year—far beyond reasonable expectations for a company that dealt largely in generic drugs, generally considered a “mature market” where high rates of competition make for modest profits.

Mylan's largest profit center was the EpiPen, a penlike device that easily injects epinephrine (adrenalin) into the skin to counteract severe allergic shock. Because each injection lasts for only a brief time, and because children at risk of allergic shock require a pen at home and at school, many families with a member at risk need to stock several pens at once. Since the medicine loses its potency after 12 to 18 months, the pens need to be replaced frequently. Mylan did not create the EpiPen: it was developed by another company that brought it to market in 1987. Mylan bought the rights in 2007; but since there was no effective competitor in the market, it had a near monopoly on epinephrine injectors.

In 2011, Mylan promoted one of its top executives, Heather Bresch, to the position of Chief Executive Officer, effective January 2012. From 2009 to 2013, the company upped the list price of a two-pen pack from \$100 to \$263; then in May 2014 (just as the new incentive system for its top executives kicked in), it doubled the price to \$461, before hiking it again in May 2015, to \$608.¹⁰

By the summer of 2016, Mylan's price gouging on this essential device—used not only by many adults but also (thanks to a marketing campaign by Mylan) by thousands of school-age children—led to a public outcry and a congressional hearing. Several senators asked the Department of Justice to investigate the company's billing practices.

How did Mylan investors fare from the company's drive to incentivize its executives to raise profits? When Bresch took charge as CEO, the stock price stood at \$22. In June 2015 it reached a high of \$73. But the public outcry against the company and the resulting congressional hearings and Justice Department investigations led the price to drop to \$36 in October 2016. The top executives' single-minded focus on hitting outsized profit metrics had led to a collapse of the company's reputation.

At the very time that Mylan's pay-for-performance scheme for top executives was bringing down the pharmaceutical company, another major corporation was being laid low by its own version of pay-for-performance. The company in question aimed not at the top of the organizational ladder but at the bottom; its incentives were not the carrots of monetary rewards but the sticks of forced termination for those who did not measure up to its performance goals.

Here's what happened. Wells Fargo, a major American bank, was functioning in a difficult economic environment. The Federal Reserve Board had lowered the rate of interest almost to the vanishing point, making it more difficult for banks to generate profits from the loans they extended. In an attempt to increase its profits, in 2011 the company encouraged "cross-selling": it set quotas for its employees to sign up customers who were interested in one of its products (say, a deposit account) for additional services, such as overdraft coverage or credit cards, which were more lucrative for the bank. Failure to reach the quota meant working additional hours without pay and the threat of termination. (Perhaps their inspiration was the Alex Baldwin character in the film *Glengarry Glen Ross*, a boss who instructs his sales force on the rules of

their sales tournament: “First prize is a Cadillac El Dorado. . . . Second prize is a set of steak knives. Third prize is you’re fired. Get the picture?”) But the quotas were set too high, given the limited number of customers who entered the bank on a daily basis. To reach their enrollment quotas, thousands of Wells Fargo bankers resorted to low-level fraud, creating PIN numbers to enroll customers in online accounts or debit cards, for example—without informing the customer. That was not the intention of the Wells Fargo management: they wanted their employees to get customers to open legitimate accounts. As it uncovered evidence of malfeasance, Wells Fargo fired some 5,300 employees for their actions. But the spate of fraud was a predictable response to the performance quotas that the company’s managers had set for their employees.

After news of the massive fraud broke in September 2016, Wells Fargo was fined \$100,000,000 by the federal Consumer Financial Protection Bureau, \$50,000,000 by the Los Angeles City Attorney, and \$35,000,000 by the Office of the Controller of the Currency. The damage to the firm was not only monetary but also reputational. The value of Wells Fargo stock fell from about \$50 in late August to \$43 by the end of September. Once again, reward and punishment for measured performance backfired.¹¹

The cases of Mylan and Wells Fargo are recent examples of an older and common pattern, by which policies of payment for measured performance lead employees to engage in actions that create long-run damage to a firm’s reputation.¹²

Is this a problem of human nature or of the propagation of the credo of paying for measured performance? To put it another way: is the notion of narrowly self-interested agents a fact of life, or is it exacerbated by a managerial ideology that

uses extrinsic rewards based upon simple models of human behavior that then become self-fulfilling prophecies? Sometimes, the way in which managers and employees are addressed by their company actually influences the way they think, so that they come to act in the narrowly self-interested way posited by the most reductive versions of principal-agent theory, with deceit and guile.¹³ In fact, it may create a situation in which the managers and employees most knowledgeable about the workings of the performance indicators are best positioned to manipulate those indicators for their own benefit, and most likely to do so.¹⁴ Take, for example, the cases of Dennis Kozlowski, the CEO of Tyco; Bernard Ebbers, CEO of WorldCom; John Rigas, CEO of Adelphia: all went to prison in the early 2000s for enriching themselves by using their detailed knowledge of their firms' transactions to manipulate the performance measures through which they were compensated.¹⁵

The reaction to these scandals led to the passage of the Sarbanes-Oxley Act of 2002, which sought to strengthen corporate accountability, in part by holding the members of the board of directors of public corporations legally liable for the accuracy of financial statements. While complying with the act has added substantial costs to the corporations, it may have strengthened public confidence in the validity of their financial reports—and so provides evidence of the advantages of transparency. But the increased legal accountability of each member of the board of directors has also imposed costs of a sort not measureable by economists. As a consultant to the boards of directors of Fortune 500 corporations (who for obvious reasons must remain nameless) told me, since the passage of Sarbanes-Oxley, board members are so focused on assuring the accuracy of the company's financial reports that they have

little time and inclination to deliberate upon the primary tasks of a board of directors, namely thinking strategically about the long-range future of the company! Thus *only* what gets measured—and potentially penalized—gets done.

THE FINANCIAL CRISIS

The financial crisis of 2008 had many causes, and some of them flowed from the attempt to substitute standardized metrics for judgment based on local knowledge, exacerbated by the effects of pay-for-performance schemes.¹⁶

As companies, including financial companies, grew larger and more diverse in their holdings, new layers of management were needed to supervise and coordinate their disparate units. From the point of view of top management, the diversity of operations meant that executives were managing assets with which they had little familiarity. That led to a search for standardized measures of performance across large and disparate organizations. Its implicit premises were these: that information which is numerically measurable is the only sort of knowledge necessary; that numerical data can substitute for other forms of inquiry; and that numerical acumen (premised upon probabilistic formulas rather than empirical research) can substitute for practical knowledge about the underlying assets.

Contributing to the financial crisis was the increasing role of financial managers who were skilled at the analysis and manipulation of metric data but did not have “concrete” knowledge or experience of the things being made or traded. As Niall Ferguson has put it, “those whom the gods want to destroy they first teach math.”

Here, in stylized form, is what happened in the lead-up to the crisis of 2008. Traditionally, banks (or individual investors)

had offered mortgages to people with whom they had direct contact. They were thus in a position to exercise judgment about who was credit-worthy and who was not. And they had an incentive to exercise that judgment: since the bank (or investor) continued to hold the mortgage, their stream of future income depended on the reliability of the mortgagee.

That began to change around the year 2000, and by 2008 that system had largely been replaced by a new one. Changes in the capital regulations of banks made the origination and holding of traditional mortgages less lucrative than holding securities comprising thousands of mortgages.¹⁷ Now the mortgages were originated not by the bank but by a mortgage brokerage firm, which made its money from the number of mortgages it processed but had no financial interest in the long-term viability of the mortgages. Mortgage originators, such as Countrywide, provided loans to people buying houses, then packaged these loans into bundles of one thousand, and sold them to a bank, such as Lehman Brothers. Since they had no long-term interest in the viability of the mortgage loans they issued, mortgage originators increasingly offered “low-doc” or “no-doc” loans, meaning that borrowers were asked to provide almost no proof that they would actually be able to repay the loans. But the bank did not hold onto them either. It created a “mortgage backed security,” an interest-bearing bond, secured by the loans, and sold these to investors. With advice from the ratings agencies (such as Moody’s), financial engineers mixed good-quality mortgages, from borrowers likely to pay, with more dubious ones, so as to squeeze the most profit out of these mortgage-backed securities,¹⁸ which they carved into “tranches” bearing different degrees of risk in return for varying rates of interest. Behind all of this was a belief in the financial sector that such diversification

was a substitute for due diligence on each asset. The idea was that if one bundled enough assets together, one didn't have to know much about the assets, or make judgments about their viability.

New, mathematically complex financial instruments were created, such as credit default swaps, which were intended to insure against the risk of sudden changes in the value of mortgage-backed securities. This was supposed to use mathematical sophistication to diminish risk, but instead led to an inability of any but a few analysts to get a clear sense of what was happening. And the creation of arcane financial instruments made effective supervision virtually impossible, both by superiors in the firm and by outside regulators.

Add to this witches' brew of dubious metrics, served up as a replacement for judgment, the fact that the remuneration of top employees at banks such as Lehman Brothers was based on pay for measured performance in the form of bonuses. Thus metrics provided the means, and pay-for-performance supplied the motivation, for undue risk-taking under conditions of opacity.¹⁹ Then, as mortgagees proved unable to make their mortgage payments, the simultaneous drop in value of mortgage-backed securities led to huge, unanticipated losses to those financial firms that had insured the securities through credit default swaps. The result was a near meltdown of the financial system.

SHORT-TERMISM

Another way in which dubious indicators of measured performance have distorted the economy is through short-termism.

Perhaps the most consequential change in the business world in recent decades has been the financialization of the

economy, above all in the United States.²⁰ As late as the 1980s, finance was an essential but limited element of the American economy. Trade in equities (the stock market) was made up of individual investors, large and small, putting their own money into stocks of companies they believed to have good long-term prospects. Investment capital was available from the major Wall Street investment banks (and their foreign counterparts), which were private partnerships in which the partners' own money was on the line. All this began to change as larger pools of capital (from pension funds, university endowments, and foreign investors) became available for investment and came to be deployed by professional money managers rather than the owners of the capital themselves. The result was a new financial system, characterized as "money manager capitalism" by the maverick economist Hyman Minsky, or "agency capitalism" by Alfred Rappaport, a business school professor.²¹

Spurred in part by these new opportunities, the traditional Wall Street investment banks transformed themselves into publicly traded corporations—that is to say, they too began to invest not just with their own funds but also with other peoples' money—and tied the bonuses of their partners and employees to annual profits. All this created a highly competitive financial system dominated by investment managers working with large pools of capital, paid for their supposed ability to outperform their peers. The structure of incentives in this environment leads fund managers to try to maximize short-term returns, and they in turn pressure the executives of the corporations whose stock they own to show gains every quarter.²²

The shrunken time horizon creates a temptation to boost immediate profits at the expense of longer-term investments,

whether in research and development or in improving the skills of the company's workforce. The emphasis placed upon quarterly earnings (which are supposed to provide transparency) and "quarterly earnings guidance"—projections by management about the firm's profitability in the coming three months—intensifies short-termism, since stock prices often rise and fall in keeping with this metric. And since the failure to reach this predicted target by the end of the next quarter may also lead to declines in stock prices, there is an inescapable temptation to game the figures so that measured performance matches the projections. It creates tremendous incentives for corporate executives to devote their creative energies to schemes that demonstrate productivity or profit by massaging the data, or by underinvesting in maintenance and human capital formation (ongoing education of employees) to boost quarterly earnings or their equivalents. The propensity for underinvestment in long-term growth is sufficiently dire that in early 2016, the CEO of the largest investment firm in the world, Larry Fink of BlackRock, wrote an open letter in which he warned, "Today's culture of quarterly earnings hysteria is totally contrary to the long-term approach that we need."²³

Gaming the metrics often takes the form of diverting resources away from their best long-term uses to achieve measured short-term goals. Take the company that, hoping to be bought out at a multiple of earnings, tries to boost its profit by laying off necessary workers. Or the CEO who smooths out corporate earnings by postponing needed investments in an effort to meet analysts' expectations for the quarter. Or the money-managers who buy shares of well-performing stocks and sell shares of underperforming stocks in time for listing in quarterly reports, disguising the fact that they bought the

high-performing stocks at high prices and that their poorly-performing stocks may have turned around had they held onto them—known in the trade as “window dressing.”²⁴

A focus on measurable performance indicators can lead managers to neglect tasks for which no clear measures of performance are available, as the organizational scholars Nelson Repenning and Rebecca Henderson have recently noted.²⁵ Unable to count intangible assets such as reputation, employee satisfaction, motivation, loyalty, trust, and cooperation, those enamored of performance metrics squeeze assets in the short term at the expense of long-term consequences. For all these reasons, reliance upon measurable metrics is conducive to short-termism, a besetting malady of contemporary American corporations.

OTHER DYSFUNCTIONS

When rewards such as pay, bonuses, and promotions are tied to meeting budget targets, there is yet another danger: distorting the information system of the organization. Managers and employees learn to lie, to massage, embellish, or disguise the numbers that are used to calculate their pay. But since these are the very numbers that executives use to coordinate the activities of the organization and decide on the allocation of future resources, the productivity and efficiency of the organization is damaged as resources are misallocated.²⁶

The attempt to substitute precise measurement for informed judgment also limits innovation, which necessarily entails guesswork and risk. As business school professors Gary Pisano and Willy Shih have argued,

Most companies are wedded to highly analytical methods for evaluating investment opportunities. Still, it remains

enormously hard to assess long-term R&D programs with quantitative techniques. . . . Usually, the data, or even reasonable estimates, are simply not available. Nonetheless, all too often these tools become the ultimate arbiter of what gets funded and what does not. So short-term projects with more predictable outcomes beat out the long-term investments needed to replenish technical and operating capabilities.²⁷

Performance metrics as a measure of accountability help to allocate blame when things go badly, but do little to encourage success,²⁸ especially when success requires imagination, innovation, and risk. Indeed, as the economist Frank Knight noted almost a century ago, entrepreneurship entails “immeasurable uncertainty,” which is not susceptible to metric calculation.²⁹

Thus, even in business and finance, metric fixation takes its toll. Businesses must be judged by more than one indicator of performance. Profit surely matters. But so, in the long run, does reputation, market share, customer satisfaction, and employee morale, which makes it possible to adapt and to find solutions to the new problems that will inevitably arise in the marketplace. In an economic world characterized by unpredictable change, there is a need for ongoing innovation, small and large, that is not readily reducible to a single performance target. Performance indicators can certainly aid, but not replace, the key functions of management: thinking ahead, judging, and deciding.³⁰

16

WHEN AND HOW TO USE METRICS

A CHECKLIST

There is nothing intrinsically pernicious about counting and measuring human performance. We all tend to project broad-ranging conclusions based on our inevitably limited experience, and measured data can serve as a useful counterpoint to those subjective judgments. The sort of measurements with which this book is concerned are performance metrics that quantify human achievement and failure. There are legitimate metrics of performance in almost every organization.

In our case studies, we've seen many instances in which metrics has been useful and effective.

In policing, computerized statistics of the incidence of crimes (Compstat) were used to good purpose, to discover where problems were greatest and where police resources were best deployed. It ran into problems only when officials used the threat of demotion or lack of promotion against those lower in the hierarchy to try to bring down the reported crime rates.

In universities, faculty evaluations can be enhanced by numerical data about publications and teaching. The metrics go awry when they are used mechanically by those who are not in a position to evaluate the accuracy and significance of the data.

In primary and secondary education, standardized tests can be used to inform teachers of how much or how little their students are learning in particular subjects. Teachers can consult with their colleagues, and adjust their methods and curriculum as a result. Problems arise when the tests become the primary basis on which teachers and schools are rewarded or punished.

In medicine, Peter Pronovost's Keystone project demonstrates how effective diagnostic metrics can be in lowering the incidence of medical errors, when what is measured accords with the professional values of practitioners. The success of the Geisinger medical system illustrates the remarkable improvements made possible by computerized measurement when integrated into an institutional culture based on cooperation, where the setting of measurement criteria and the evaluation of performance are done by teams that include physicians as well as administrators. In both cases, metrics were used in ways that appealed to intrinsic motivation and to professionalism. But elsewhere in the medical system, as we've seen, the use of reward for measured performance sometimes proved fruitless or led to perverse outcomes.

Reflections on the best use of performance metrics by the U.S. Army in its counterinsurgency campaigns showed that while standardized metrics are often deceptive, metrics developed to fit the specific case, especially by practitioners with local experience, could be genuinely informative. The challenge in such cases is to abandon universal templates and discover what is worth counting, and what the numbers actually mean in their local context.

As we've seen time and again, measurement is not an alternative to judgment: measurement *demand*s judgment: judgment about whether to measure, what to measure, how

to evaluate the significance of what's been measured, whether rewards and penalties will be attached to the results, and to whom to make the measurements available.

Should you find yourself in a position to set policy, here are the questions you should ask, and the factors you should keep in mind, in considering *whether* to use measured performance, and if so, *how* to use it. They constitute a checklist of successful performance measurement. Given what we've said about the hazards of metric fixation, consider at every point that the best use of metrics may be not to use it at all.

THE CHECKLIST

1. What *kind* of information are you thinking of measuring? The more the object to be measured resembles inanimate matter, the more likely it is to be measureable: that is why measurement is indispensable in the natural sciences and in engineering. When the objects to be measured are influenced by the process of measurement, measurement becomes less reliable. Measurement becomes much less reliable the more its object is human activity, since the objects—people—are self-conscious, and are capable of reacting to the process of being measured. And if rewards and punishments are involved, they are more likely to react in a way that skews the measurement's validity. By contrast, the more they agree with the goals of those rewards, the more likely they are to react in a way that enhances the measurement's validity.
2. How *useful* is the information? Always begin by reminding yourself that the fact that some activity is measureable does not make it worth measuring, indeed, the ease of measuring may be inversely proportional to the significance of what is measured. To put it another way, ask your-

self, is what you are measuring a proxy for what you really want to know? If the information is not very useful or not a good proxy for what you're really aiming at, you're probably better off not measuring it.

3. How useful are *more* metrics? Remember that measured performance, when useful, is more effective in identifying outliers, especially poor performers or true misconduct. It is likely to be less useful in distinguishing between those in the middle or near the top of the ladder of performance. Plus, the more you measure, the greater the likelihood that the marginal costs of measuring will exceed the benefits. So, the fact that metrics is helpful doesn't mean that more metrics is more helpful.
4. What are the costs of *not* relying upon standardized measurement? Are there other sources of information about performance, based on the judgment and experience of clients, patients, or parents of students? In a school setting, for example, the degree to which parents request a particular teacher for their children is probably a useful indicator that the teacher is doing something right, whether or not the results show up on standardized tests. In the case of charities, it may be most useful to allow the beneficiaries to judge the results.
5. To what *purposes* will the measurement be put, or to put it another way, to *whom* will the information be made transparent? Here a key distinction is between data to be used for purposes of internal monitoring of performance by the practitioners themselves versus data to be used by external parties for reward and punishment. For example, is crime data being used to discover where the police ought to deploy more squad cars or to decide whether the precinct commander will get a promotion? Or is a surgical

team using data to discover which procedures have worked best or are administrators using that same data to decide whether the hospital will be financially rewarded or penalized for its scores? Measurement instruments, such as tests, are invaluable, but they are most useful for internal analysis by practitioners rather than for external evaluation by public audiences who may fail to understand their limits. Such measurement can be used to inform practitioners of their performance relative to their peers, offering recognition to those who have excelled and offering assistance to those who have fallen behind. To the extent that they are used to determine continuing employment and pay, they will be subject to gaming the statistics or to outright fraud.

Remember that, as we've seen, performance metrics that link reward and punishment may actually help reinforce intrinsic motivation when the goals to be rewarded accord with the professional goals of the practitioners.¹ If, on the other hand, the scheme of reward and punishment is meant to elicit behavior that the practitioners consider useless or harmful, the metrics are more likely to be manipulated in the many ways we've explored. And if the practitioners are too geared toward extrinsic reward, they may well react by focusing their activity on what is measured and rewarded, at the expense of other facets of their work that may be equally important. For all these reasons, "low stakes" metrics are often more effective than when the stakes are higher.

Recall that direct pay-for-performance works best to the degree that people are motivated by extrinsic reward rather than intrinsic motivation, that is, when they care about making more money rather than about the other potential benefits of their work, social and intellectual.

That may be because they are in a field, such as finance, in which people measure their own vocational success almost entirely in terms of the amount they earn. (As we've noted, that doesn't preclude them from using their earnings for a wide range of purposes, including selfless ones.) It is when the job offers few other attractions—when it is repetitious and leaves little room for the exercise of choice, for example replacing windshields or preparing hamburgers—that pay for measured performance is more likely to work.

6. What are the *costs* of acquiring the metrics? Information is never free, and often it is expensive in ways that rarely occur to those who demand more of it. Collecting data, processing it, analyzing it—all of these take time, and their expense is in the opportunity costs of the time put into them. To put it another way, every moment you or your colleagues or employees are devoting to the production of metrics is time not devoted to the activities being measured. If you're a data analyst, of course, producing metrics *is* your primary activity. For everyone else, it's a distraction. So, even if the performance measurements are worth having, their worth may be less than the costs of obtaining them. Remember, too, that those costs in human time and effort are themselves almost impossible to calculate—another reason to err on the side of caution.
7. Ask why the people at the top of the organization are demanding performance metrics. As we've noted, the demand for performance measures sometimes flows from the ignorance of executives about the institutions they've been hired to manage, and that ignorance is often a result of parachuting into an organization with which one has little experience. Since experience and local knowledge

matter, lean toward hiring from within. Even if there is someone smarter and more successful elsewhere, their lack of particular knowledge of your company, university, government agency, or other organization may not outweigh the benefits of hiring from within.

8. *How and by whom* are the measures of performance developed? Accountability metrics are less likely to be effective when they are imposed from above, using standardized formulas developed by those far from active engagement with the activity being measured. Measurements are more likely to be meaningful when they are developed from the bottom up, with input from teachers, nurses, and the cop on the beat. That means asking those with the tacit knowledge that comes from direct experience to provide suggestions about how to develop appropriate performance standards.² Try to involve a representative group of those who will have a stake in the outcomes.³ In the best of cases, they should continue to be part of the process of evaluating the measured data.

Remember that a system of measured performance will work to the extent that the people being measured believe in its worth. So far, in this chapter, we've taken the perspective of those in a position to decide whether and how to institute metrics. But what if you are not in such a position, if you're further down in the organizational hierarchy, where you are expected to execute metrics—a mid-level manager, say, or the head of an academic department? Then, you face a choice. If you believe in the goals for which the information is being collected, then your challenge is to provide accurate data in the most efficient way possible, one that demands the least time of you and those you manage. If, by contrast, you believe that the goals

are dubious and the process wasteful, you might try to convince your superiors of that (perhaps by giving them a copy of this book). If that fails, then your task is to provide data in a way that takes the least time, meets minimal standards of acceptability, and won't harm your unit.

If you're near the *top* of the organization, making decisions about metrics, reread the previous paragraph, keeping in mind the different ways in which those below you might react. Metrics works best when those measured buy into its purposes and validity.⁴

9. Remember that *even the best measures are subject to corruption or goal diversion*. Insofar as individuals are agents out to maximize their own interests, there are inevitable drawbacks to all schemes of measured reward. If, as is currently still the case, doctors are remunerated based on the procedures they perform, that creates an incentive for them to perform too many procedures that have high costs but produce low benefits. But pay doctors based on the number of patients they see, and they have an incentive to see as many patients as possible, and to skimp on procedures that are time-consuming but potentially useful. Compensate them based on successful patient outcomes, and they are more likely to cream, avoiding the most problematic patients.⁵

That doesn't mean that performance measures should be abandoned just because they have some negative outcomes. Such metrics may still be worth using, despite their anticipatable problems: it's a matter of trade-offs. And that too is a matter of judgment.

10. Remember that sometimes, recognizing the limits of the possible is the beginning of wisdom. Not all problems are soluble, and even fewer are soluble by metrics. It's not true

that everything can be improved by measurement, or that everything that can be measured can be improved. Nor is making a problem more transparent necessarily a step to its solution. Transparency may make a troubling situation more salient, without making it more soluble.

In the end, there is no silver bullet, no substitute for actually knowing one's subject and one's organization, which is partly a matter of experience and partly a matter of unquantifiable skill. Many matters of importance are too subject to judgment and interpretation to be solved by standardized metrics. Ultimately, the issue is not one of metrics versus judgment, but metrics as informing judgment, which includes knowing how much weight to give to metrics, recognizing their characteristic distortions, and appreciating what can't be measured. In recent decades, too many politicians, business leaders, policymakers, and academic officials have lost sight of that.