Faculty Research Working Papers Series

Performance Improvement and Performance Dysfunction: An Empirical Examination of Impacts of the Emergency Room Wait-Time Target in the English National Health Service

Steven Kelman
John F. Kennedy School of Government – Harvard University

John N. Friedman
University of California, Berkeley

August 2007

RWP07-034

PERFORMANCE IMPROVEMENT AND PERFORMANCE DYSFUNCTION: AN EMPIRICAL EXAMINATION OF IMPACTS OF THE EMERGENCY ROOM WAIT-TIME TARGET IN THE ENGLISH NATIONAL HEALTH SERVICE

Steven Kelman

Harvard University

John N. Friedman

University of California, Berkeley

ABSTRACT

The literature on the use of performance measurement in government has featured prominent attention to hypothesized unintended dysfunctional consequences such measurement may produce. We conceptualize these dysfunctional consequences as involving either effort substitution (reducing effort on non-measured performance dimensions) or gaming (making performance on the measured performance dimension appear better, when in fact it is not). In this paper, we examine both performance impacts and dysfunctional consequences of establishment in the British National Health Service of a performance target that no patient presenting in a hospital accident and emergency department (emergency room) wait more than four hours for treatment. Using data from all 155 hospitals in England, we find dramatic wait-time performance improvements between 2003 and 2006, and no evidence for any of the dysfunctional effects that have been hypothesized in connection with this target. We conclude by discussing when one would expect dysfunctional effects to appear and when not.

KEY WORDS: PERFORMANCE MEASUREMENT, HOSPITAL PERFORMANCE, EMERGENCY ROOM PERFORMANCE

During the past decades, there has been a significant increase in use of non-financial performance measurement in government as a tool to improve both democratic accountability and organizational performance (Cave, Cogan, and Smith, 1990; Carter, Klein, and Day, 1992; Hatry, 1999; Heinrich, 2003; Talbot, 2005). Examples of such performance measures include anything from crime rates for the police to average wait time at a Division of Motor Vehicles office to the percentage of citizens calling a Social Security call center who are satisfied with the service they received. One may see non-financial performance measures as the public sector's counterpart to profit as a performance measure for firms. A new attention to performance measurement appeared with force in the U.K. starting during the Thatcher government of the 1980's and continuing under Blair's "New Labour" after 1997. It emerged strongly in the U.S. with passage of the Government Performance and Results Act in 1993.

Setting targets for improved performance may increase organizational performance through several routes. First, a large psychological literature establishes that giving people goals motivates better performance, especially if attached to incentives for goal achievement, but, as long as people accept the goals, even without incentives (e.g. Locke and Latham, 1990a, 1990b, 2002). Second, setting a performance target for one endeavor rather than another sends people a signal about, of the many possible activities on which employees could focus on the job, on which their bosses want them to focus--"what gets measured gets noticed," to use the common phrase. Third, performance information improves learning within a unit by providing a source of feedback about the

We would like gratefully to acknowledge the assistance of Stephen Morris, Lis Nixon, Liz Chart, Adrian Masters, Duncan Selbie, and Bill Alexander in helping us locate data used in this paper and in answering a large number of questions from us, and of Paul Corrigan in arranging access to the U.K. Department of Health. We also wish to acknowledge funding support from NIA Grant T32-AG00186 and the Robert Wood Johnson Foundation.

success of previous endeavors -- imagine how much harder it would be to learn to throw darts if one didn't get feedback about where the dart one had previously thrown had landed – and across units by allowing learning from similar units that are more successful at their tasks (Ilgen, Fisher & Taylor, 1979; Huber, 1991; Hedlund, 1994; Argote, 1999; Metzenbaum, 2003; Kelman, 2006).

However, scholarly writing on performance measurement in government has long featured concern about dysfunctional reactions to performance measurement – in fact, it has often been as preoccupied, or even more so, with dysfunctional as with functional responses. These worries, particularly in the public management literature, evoke the spirit, and sometimes the letter, of Merton's (1936) idea of the "unintended consequences of purposive social action." One paper on performance measurement (Smith, 1995) is in fact straightforwardly titled "On the Unintended Consequences of Publishing Performance Data in the Public Sector," and Radin's (2006) Challenging the Performance Movement analysis of performance measurement begins with a discussion, complete with cite to Merton, of this theme. An early chapter in deBruijn's (2007) Managing Performance in the Public Sector is entitled, "Perverse Effects of Performance Measurement." (See also Grizzle, 2002; van Thiel and Leeuw, 2002.)

Hirschman (1991), presenting this as an example of "the rhetoric of reaction," refers to this as "the perversity thesis."

> It is not just asserted that a movement or a policy will fall short of its goal or will occasion unexpected costs or negative side effects: rather, so goes the argument, the attempt to push society in a certain direction will result in its moving…in the opposite direction. …In current debates, it is often invoked as the counterintuitive, counterproductive…effect of some… "well-intentioned" public policy. Attempts to reach for liberty will make society sink into slavery…, and social welfare programs will create more,

> rather than less, poverty. Everything backfires. (pp. 11-12, emphases in original)

And, indeed, the spirit of these criticisms does suggest a sense of futility about the ability of purposive action to improve government performance.[1]

Concerns with dysfunctional responses to performance measurement go back to the earliest discussions of the topic in organization theory. Indeed, the very first two issues of Administrative Science Quarterly, in 1956, featured papers on this problem. One (Berliner, 1956), entitled "A Problem in Soviet Business Management," identified the phenomenon of "storming," whereby Soviet firms rushed at the end of the month to meet monthly production quotas, creating quality and equipment maintenance problems. (Ever since, some critics – e.g. Meyer and Gupta, 1994: 361-62; Smith, 1995; Bevan and Hood, 2006 -- have seen analogies between problems of performance measurement for government programs and those created by their use in Soviet planning.) The second early ASQ paper (Ridgeway, 1956) was actually titled "Dysfunctional Consequences of Performance Measurements" and discussed many of the problems that have received frequent attention since. Blau's organization studies classic, The Dynamics of Bureaucracy (1955), appearing around the same time, took up this problem as well in the context of a state employment agency. While generally positive towards the impacts of measuring performance, Blau also devotes a section (pp. 40-44) to "dysfunctional consequences," such as deciding on the order one worked on cases based on monthly case quotas rather than the cases' logical priority, or asking clients being temporarily laid off to enter the system so they could be measured as both a job opening and a placement.

[1] As Hirschman notes (1991: 38), Merton had originally wrote about "unanticipated" consequences, and not assumed these were necessarily negative.

In this paper we explore responses to a performance target during the Blair government in the United Kingdom for wait times in English[2] hospital "Accident and Emergency" (A&E) departments -- equivalent to emergency rooms in the United States – run by the governmental National Health Service. The paper has four aims. First, we discuss theoretically – using literature not only from public management but also economics, organization theory, and accounting – why one might expect dysfunctional responses to the adoption of performance measures in an organization and what the different categories of such distortions might be. We illustrate this with examples of distortions predicted for the English A&E wait time performance target. Second, we present empirical results, based on econometric analysis of data from all English hospitals during the period 2003-06, on both performance improvements in wait times and on presence of the predicted dysfunctional effects. We find that waiting-time performance improvement was dramatic and that dysfunctional responses, as far as we can tell, entirely absent. Indeed, in a number of cases, the sign of statistically significant effects predicted by those worried about dysfunctional effects went in the "wrong" direction, i.e. that better waiting-time performance was associated with a lower level of problems predicted by a dysfunctional effects story. Third, we discuss why the predicted distortionary effects failed to appear in this instance, and, through that discussion, present limitations and cautions about the dysfunctional effects story for performance measurement more generally. Fourth, we note when it would promote overall organizational performance to adopt performance measurement regimes, despite the possible presence of distortionary effects.

---

[2] This and other health targets applied only to English hospitals, not to regionally devolved parts of the NHS in Wales, Scotland, or Northern Ireland.

BACKGROUND

England's National Health Service (NHS) was established in 1948 as the nation's primary healthcare system.[3] Reform of the NHS, and more generally of public service provision, was a key element of Tony Blair's election platform in 1997. In contrast to the movement towards privatization under Margaret Thatcher and John Major, Blair proposed an aggressive program to achieve performance improvement using performance measurement standards, or "targets" (Kelman, 2006). One key target established for the NHS was a reduction to four hours of the maximum time a patient was required to wait for treatment[4] in an A&E department. This target responded to the fact that a leading source of citizen dissatisfaction with the NHS was length of wait times. A 2000 NHS Report (Department of Health, 2000) set an interim target of 90% compliance with the A&E target by March 2003 and an eventual goal of 100%.

In January 2003, responding to an apparent lack of attention to, or progress on, attaining the A&E wait time target, the Department of Health, which oversees the NHS, announced an increased level of attention to this target. First, the Department announced that A&E wait times relative to the interim target would be included, for the first time, in a "star rating" system for hospitals. "Star ratings," prepared by an independent government audit body, give hospitals overall scores between zero and three stars, and are one of a number of "league tables" measuring comparative performance of public organizations that have been established in the U.K.[5] The first such ratings for hospitals

---

[3] This section closely follows Friedman and Kelman (2007).

[4] Or for admission into an inpatient ward.

[5] Star ratings have also been established for schools and for local governments. Other measured areas in the hospital star ratings included the wait time for elective surgery, the death rate following major operations, survey feedback from patients and doctors, as well as a number of softer criteria such as a consultant appraisal and the quality of hospital food.

were prepared in 2001, and they became required by law in 2002. A&E performance would be measured for the 2003 star ratings during the final week of March.

A year later, in January 2004, the Prime Minister's Delivery Unit, an organization Blair created to work on performance targets, released a "5-Point Plan" for meeting the A&E target. This document announced monetary incentives for hospitals that met the target, along with weekly monitoring of A&E performance by the Delivery Unit and consulting support for hospitals that were falling far short of meeting the target. In each of the next five fiscal quarters, hospitals would receive a lump-sum grant of £100,000 if the percent of patients treated within four hours across the entire quarter rose above a threshold. The thresholds began at 94% for March 2004 and increased by a single percentage point each quarter thereafter until reaching 98% for the first quarter of 2005.[6] The report explicitly stated that incentive payments would end after March 2005.

By the end of the period we measured (mid-September 2006) 76% of English hospitals were meeting the four-hour waiting target, compared with 1% at in January 2003, when central government attention to this issue began.[7] This improvement occurred in connection with improvements during the two incentive and special central government attention periods discussed above.

However, for many of the Blair government's performance targets, there has been an undertone – perhaps better characterized as a roar – of teeth-gnashing about dysfunctional responses to the efforts. These dysfunctional responses, it has been

---

[6] The 5-Point Plan changed the final target from 100% to 98% of patients handled within four hours, based on consultations with doctors about justifiable exceptions to the four-hour treatment standard.

[7] In both cases, these figures are for the mean percentage of hospitals reaching the target, calculated for each week of the period.

suggested, vitiate or even destroy the significance of reported performance improvements such as those noted above.

One gets a flavor for these concerns by examining the first major effort undertaken to encourage attainment of the A&E target, the inclusion of A&E waiting-time performance in 2003 hospital star ratings. As noted earlier, performance for the purpose of the star rating was measured during one week, announced well in advance; elsewhere (Friedman and Kelman, 2007) we have referred to this measurement week as a "sweeps week" for the hospitals.[8]

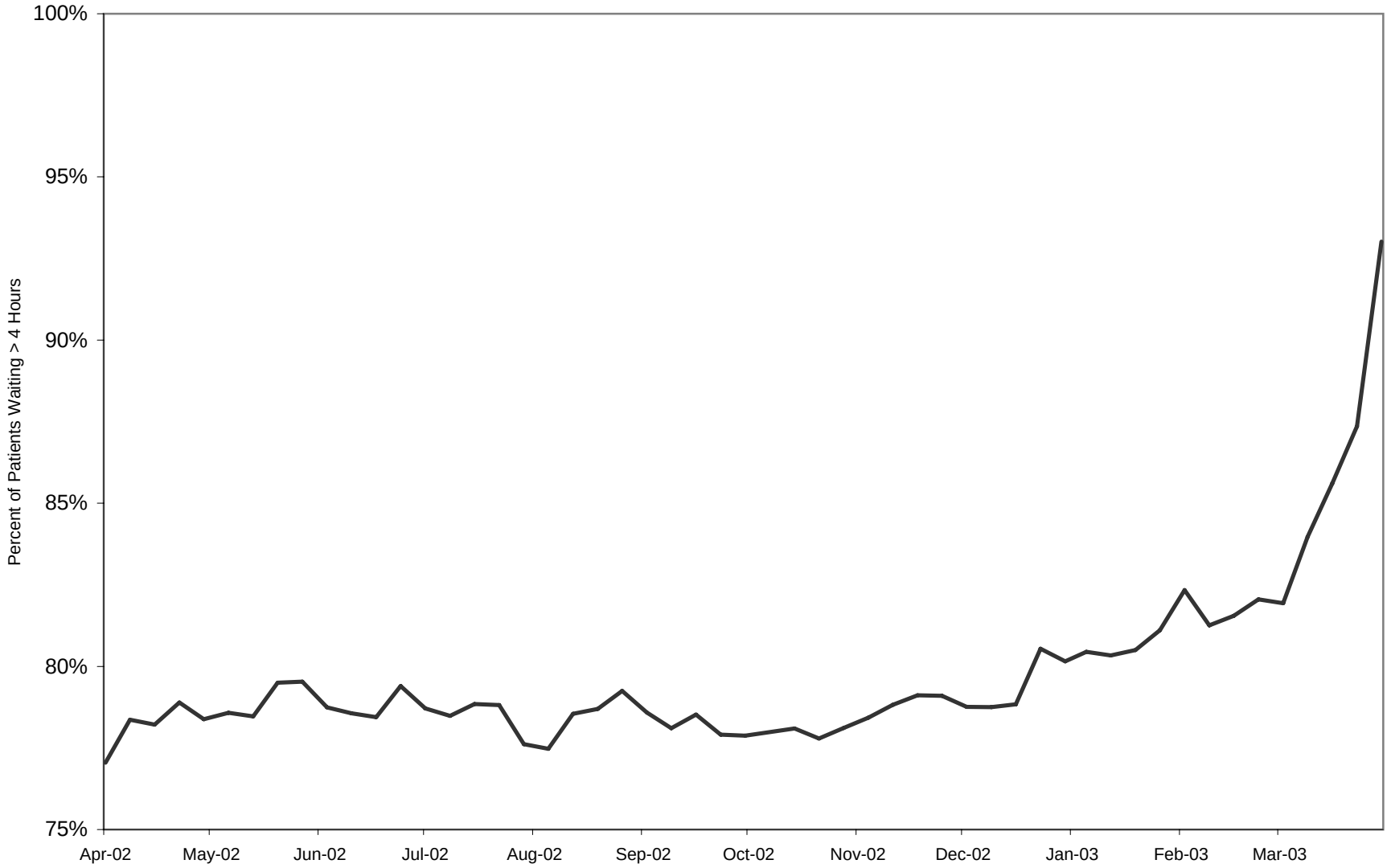
During that week, A&E waiting-time performance spiked up dramatically from that during previous weeks and months, attaining average levels well above those ever registered before. This is seen in Figure 1, which shows average waiting-time performance for all hospitals in England in the weeks before and during the one-week measurement period. The month before "sweeps week," the mean percentage of patients treated within four hours was 85%. During "sweeps week," the mean was 93%.

[FIGURE 1 GOES ABOUT HERE]

Contemporaneous media coverage in the U.K. was quick to notice and criticize this sudden spike. The week the measurement was taking place, The Guardian (Meikle, 2003) ran the headline, "Wait times in A&E 'Fiddled.'" The article reported that "according to allegations made anonymously to the Guardian this week," some hospitals had "employed extra staff, paid existing staff for more shifts, and delayed routine operations to free beds so they can admit emergency patients more quickly." The

[8] Four times per year, a week of television programming is designated a "sweeps week." Program ratings from viewer diaries recorded during this week are then used to set advertising rates until the next measurement week.

Figure 1: Emergency Room Waiting Times



Financial Times (Timmins, 2003) ran a similar story that week under the headline, "Hospitals Make Frantic Efforts to Hit A&E Targets." A few months later, a British Medical Association survey showing special efforts made during "sweeps week" received significant media coverage as well. The Guardian (Carvel, 2003) reported:

> Two-thirds of NHS accident and emergency department in England faked improvements in their wait times during the week chosen by ministers to measure their performance, a survey by the British Medical Association showed yesterday. It accused the government of conniving in the scam and using "immoral" tactics to persuade the public that it was achieving its political targets. Most consultants thought wrong clinical decisions were made as a result of the exercise.
> The association's survey of 500 A&E consultants found that 56% hired extra doctors and nurses for the seven days in March when they knew they would be assessed on how well they were meeting the government target to keep maximum waits below four hours. A quarter required staff to work double or extend shifts and 14% cancelled routine surgery to free beds to relive bottlenecks in A&E.
> …Don Mackechnie, the chairman of the BMA's A&E committee said: "I am appalled to see how A&E departments have been forced into taking extraordinary measures for a week-long period just to meet political targets. It is completely immoral for the government to claim that it is raising the standard of performance in the NHS when this is how they measure it."

The Sunday Times (Carr-Brown and Tonner, 2003) reported that "(h)ospitals are cheating NHS performance league tables by putting extra staff into casualty departments during the only week of the year when they are assessed." The article continued, "There are genuine concerns that patient care may be jeopardised by attempts to achieve this four-hour target." This event has also been noted in scholarly treatments of performance measurement in the public sector (Bevan and Hood, 2006; deBruijn, 2007).

DISTORTIONS PERFORMANCE MEASURES CAN PRODUCE

In this section we discuss two possible distortions arising from use of performance measures: (1) effort substitution and (2) gaming.

Effort Substitution

An organization's (or an individual's) performance typically has more than one dimension. Widget production performance includes both the quantity and the quality of widgets produced. The U.S. Forest Service has multiple goals, such as attaining economic value from exploitation of timber resources, providing recreational opportunities for users, and protecting wilderness resources. The focusing function of performance measures – "what gets measured gets noticed" – can have a dark side – "what gets measured gets noticed." If performance measures leave out importantly relevant aspects of an organization's performance, then "measured performance differs from total contribution" (Gibbons, 1998: 120).[9]

This problem is given formal expression in a classic paper by Holmstrom and Milgrom (1991) that sought to explain why many employment contracts paid employees a fixed wage rather than piece rates tied to performance measures. The paper argued (1991: 25, emphasis in original) that this phenomenon can be seen as a response to a situation where "the principal either has several different tasks for the agent or agents to perform, or the agent's single task has several dimensions to it" – say, responsibility for producing both high volume and good quality output.[10] Noting that performance measurement "serves to direct the allocation of the agents' attention among their various duties," Holmstrom and Milgrom argue that "if volume of output is easy to measure but the quality is not, then a system of piece rates for output may lead agents to increase the

[9] This may occur either because of a conscious or unreflective decision to leave out other performance measures, or because the unmeasured elements of performance are simply difficult to measure. The latter problem is captured in the aphorism (Talbot, 2005: 503) that "we make important what can be measured, because we cannot measure what is important."

[10] Kerr (1976: 779-80) had made a similar argument 15 years earlier, but it appeared in The Academy of Management Journal rather than an economics journal, and was little-noted by economists.

volume of output at the expense of quality."[11] Since Holmstrom and Milgrom, in economics this problem has generally been named "multitasking" or "effort substitution." The problem is especially serious when, from the principal's perspective, the two tasks are both necessary parts of a whole – it may be worthless, from the principal's perspective, for the agent to increase the quantity of production if ignoring quality makes the production unsaleable.

Blau (1955) noted this problem early on in the literature when he pointed out that measuring employment agency counselors only by the number of interviews they conducted would discourage them from devoting time to help clients find jobs. Heinrich (1999) found that attention in job training programs to a cost-per-placement performance measure had a negative impact on service quality. In educational testing, this problem would appear in the form of reduction in instruction in social studies or foreign languages (which typically are not subject to statewide standardized testing) to make way for increased instruction in reading and math (which are). This problem may be seen as likely to be especially important in the public sector because of the larger number of goals public organizations are often asked to pursue, and arguably because in the public sector more goals are, empirically, difficult to quantify, or involve improvements whose effects appear only after many years (Rainey, 1993; Smith, 1995).

Effort substitution involves goals that may vary in terms of how closely they are to each other. Some who complain about the effects of standardized school tests worry that testing reading and math skills reduces school resources devoted to history or

[11] Similar theoretical points, albeit with less formal elegance, appear in the public administration and political science literatures, e.g. Wilson, 1989; Smith, 1995; Kravchuk and Schack,1996; Bohte and Meier, 2000; Bevan and Hood, 2006. Wilson sees this as a form of Gresham's Law, where measured behavior drives out unmeasured behavior. Smith calls the phenomenon "tunnel vision." Bevan and Hood call it "output distortion."

athletics; here the effort substitution involves goals that are quite distinct. Others complain that multiple-choice standardized tests for reading skills test certain kinds of reading knowledge (say of vocabulary words or the ability to understand the contents of a text) at the expense of, say, the ability creatively or expressively to use language (Radin, 2006); here, effort substitution involves goals that are closer to each other, both dimensions of providing students with language skills. Cream-skimming in job-training programs (abandoning hard-to-place cases and attending only to easier ones; see e.g. Heckman, Heinrich, & Smith, 2002) also involves effort substitution among goals that are closer to each other.

The more peripheral the aspect of performance that a measure captures, the more that effort substitution in the direction of that measure may be seen as outright effort misdirection -- performance improvement on behalf of a measure that itself inappropriately specifies the underlying goal the measure is designed to represent. Air Force General John Jumper, complained, when he took over the Air Combat Command, "We once had a quality Air Force that was ruined by a program called 'Quality Air Force'" (Jumper, 2000). In saying this, he was arguing that the metrics the Air Force was using to measure quality were so peripheral to what truly constituted quality that directing effort towards attaining those performance measures had a negative overall effect on the goal of achieving quality. Similarly, Heckman, Heinrich, and Smith (2002) find that long-run earnings increases for workers in job-training programs (presumably the goal the program seeks) correlate only weakly with the short-run increases the program measures; this may be seen as an instance of the same phenomenon, whereby the measure used was only distantly related to the underlying goal) being sought.

For A&E departments, the most dramatic example of effort substitution would be to improve service speed (wait time) at the expense of the quality of care the patient receives. This apprehension lay at the heart of concerns expressed in British press accounts that the A&E wait time target threatened to produce poor "clinical decisions" (see also Bevan and Hood, 2006). At the extreme, patients could die from receiving care whose quality suffered from rushing the patient through the system. Alternatively, poor-quality care would presumably produce more return visits to the A&E department, since the patient had not gotten his or her problem properly treated the first time.

Effort substitution might also occur in a quite direct way by transferring resources such as doctors and nurses from non-emergency activities in other parts of the hospital (such as inpatient elective surgery) into the A&E department so that A&E could better meet its wait time target (Bevan and Hood, 2006). If this happened, improved A&E wait time performance would occur at the expense of reduced performance elsewhere.

A third example of effort substitution would be redistribution of wait times -- increasing short patient waits, or perhaps at the extreme even increasing the average wait, in order to meet the target, since this was expressed as a four-hour threshold. In other words, a hospital might keep patients who otherwise would have been treated in 30 minutes waiting for nearly four hours, in order to devote attention to patients who would otherwise breach the four-hour threshold (Smith, 1995[12]; Bevan and Hood, 2006). To the extent wait and treatment time in an A&E is determined by the severity of a patient's situation (with sicker patients treated faster), if there is redistribution of wait times in A&E departments away from treating patients quickly to getting them treated in under four hours, such redistribution would entail an increase in wait times for more serious

[12] He discusses this issue in the context of wait times for elective surgery in inpatient wards.

patients in exchange for a decrease for less serious patients. Unless the marginal cost to patients of waiting is rapidly increasing, such effort substitution, all else equal, would also lower the quality of care, especially when serious patients have life-threatening conditions requiring rapid attention.[13]

The presence of effort substitution occurring in response to the A&E target can be tested through the following hypotheses:

Hypothesis 1: Better performance in meeting the English A&E wait time target is associated with lower-quality care for patients presenting in the A&E department.

Hypothesis 2: Better performance in meeting the English A&E wait time target is associated with substitution of resources from elective activities in other parts of the hospital into the A&E department.

Hypothesis 3: Better performance in meeting the English A&E wait time target is associated with a decrease in the percentage of patients treated within two hours and an increase in mean wait time.

Gaming

Effort substitution produces performance improvement along the dimension being measured – so the quantity of output does indeed increase, even as its quality declines. By contrast, behavior that consumes real resources but produces no genuine performance improvement even on a measured dimension may be referred to as "gaming."[14] Gaming creates clear net social costs, both because of the resources it consumes (in this sense similar to rent-seeking behavior) and because it may lead to less-efficient production.

---

[13] Such reduced quality of care would presumably also affect death and revisit rates.

[14] This definition of gaming is similar to Baker 1992: 600. deBruijn (2007: 19) refers to situations where "(t)he performance on paper has no social significance," which is an intuitive way of stating what we mean by gaming. By contrast, Bevan and Hood (2006) use the term "gaming:" to refer to both what we call effort substitution and gaming.

The limiting case of gaming is outright data falsification or cheating. In the laboratory, Schweitzer, Ordonez, & Douma (2004) found that, when offered a reward for meeting a goal for a word-puzzle game, somewhat fewer than one in eight subjects claimed falsely to have met the goal, under conditions that led the subjects to believe they would not be caught . Jacob and Levitt (2003) detected real-world evidence for cheating in Chicago school standardized test results by examining, for example, unexpected test score fluctuations among students across years (students whose performance goes up dramatically from one grade to the next and then falls back the following year) and high variance within a class in the correlations among student responses to different questions in the exam (suggesting answers to some questions were tampered with).

But there are many other examples of gaming that fall short of outright data falsification. An airline or train service may improve its "on-time" punctuality performance by "increasing the predicted length of a flight," thus improving "their official statistics without actually improving their performance" (Gormley and Weimer, 1999: 149). There is an extensive literature in accounting on "earnings management," activities by a firm to adapt reported earnings for a given period either upwards or downwards, taking advantage of discretionary features in accounting standards such as provisions for bad debts. Earnings management occurs to influence either contractual outcomes such as incentive compensation tied to reported earnings or financial market perceptions tied to the smoothness of earnings progression over time (Healy, 1985; McNichols and Wilson, 1989; Healy and Wahlen, 1999). A review of the literature (Healy and Wahlen, 1999) noted evidence, for example, that firms overstated earnings prior to new equity offerings or to meet financial analysts' earnings expectations, and

deferred earnings to a future period if manager earnings targets for bonuses were unlikely to be met anyway. Alternatively, product shipment dates may be manipulated to fit into a certain month or quarter: Jensen (2003) recounts the example of unfinished products being shipped to another country so the sale could be recognized earlier. Courty and Marschke (2004) found evidence of similar timing adjustments for graduation from job training in training programs that measure job-placement performance. There is also empirical evidence from Texas and Florida that school districts exclude some students likely to perform poorly in standardized tests from the pool of those taking the test by placing them in special education or subjecting them to long-term expulsion, both removing the obligation to have the child tested (Bohte and Meier, 2000; Figlio, 2005; Cullen and Reback, 2006).

Gaming uses real resources (spent on organizing the manipulation), and may also lower performance through inappropriate decisions. In Jensen's example about shipping product to move sales into an earlier reporting period, the shipping to a distant location created unnecessary transport costs. Artificially moving production across periods might also produce unnecessary overtime costs. Courty and Marschke's job-training research found that gaming sometimes led to inappropriate truncation of training, which in turn produced lower earnings gains for trainees. Removing pupils from a testing pool may have a negative impact on the performance of those children if they are inappropriately put into special education or expelled, compared with a counterfactual where those children were kept in regular classes (either because the test was not gamed or because there was no test at all). Finally, if gaming creates an incorrect impression of acceptable performance, the performance-promoting pressures that performance measurement

produces will be short-circuited, compared with a counterfactual where the test was not gamed.

The boundary between effort substitution and gaming can be imprecise. Since a firm's owners care about the net present value of the total stream of the firm's earnings, earnings manipulation that moves a given quantum of earnings forward in time may be seen as increasing the firm's overall value -- effort substitution -- while one delaying earnings would reduce it --gaming (Dechow and Skinner, 2000). In the context of "cream-skimming" in job-training programs, if it is empirically the case that the job training organization makes no contribution to a "cream-skimmed" jobseeker's prospects and simply claims credit for successes that would have occurred anyway, this constitutes gaming, because performance is not improved on any dimension, and effort is spent on non-value added activity (working with the easy-to-place jobseeker) that could have produced some improvements for the hard-to-place. By contrast, say that the training organizations do help easier-to-place jobseekers, and at modest cost compared to the effort it would take to help harder-to-place ones – illustratively, it may take one unit of organizational effort to prepare the easy-to-place for a job and three units for the hard-to-place. Winter (2005), for example, found that the more refugee caseworkers in Denmark engaged in cream-skimming (at least in difficult task environments with many refugees), the shorter the average time it took for refugees to find employment. In this case, creaming constitutes effort displacement from aiding hard-to-place workers to aiding easy-to-place ones; whether this focus is socially justified is a value question.[15] In the case of television "sweeps week," networks attempt to schedule the best, most

[15] This analysis is similar to Heckman, Heinrich, and Smith's (2002) discussion of the efficiency consequences of creaming.

outrageous, or most eye-catching episodes of a given series, or blockbuster movies, during these weeks; if episodes of this appeal would otherwise not have been shown at all, "sweeps week" may be seen as involving effort substitution, while if the timing of programming that would have been shown anyway is merely moved into sweeps week, this constitutes gaming (producing a net social loss if viewers would have rather seen the programs, say, around Christmas or Thanksgiving).

For the A&E target, the performance improvement during "sweeps week" in March 2003 would be seen as an example of gaming almost identical to the earnings manipulation discussed in the accounting literature.

Furthermore, when patients requiring serious treatment come into the A&E department (say with a heart attack), they clearly cannot be fully treated within four hours, so the target specifies that they must be admitted to inpatient wards, where further treatment will take place, within four hours. It has been suggested (Bevan and Hood, 2006) that patients hitting up against the four-hour treatment target would be taken out of A&E and admitted into hospital inpatient wards. This would also be an example of gaming, costly since the cost of care in inpatient wards is higher and because free beds are often in short supply.

A final suggestion (Bevan and Hood, 2006; deBruijn, 2007) has been that gaming occurred at the front end of the process -- that ambulances arriving at A&E departments would wait outside, keeping the sick patient in the ambulance, if the A&E was crowded and didn't want the patient's clock to start ticking (since the patient in the ambulance had not entered the A&E system), again a response that would worsen performance because

at a minimum it would make the patient suffer more inside an ambulance and in the worst case would hurt treatment quality if keeping them in an ambulance delayed needed care.

The presence of effort substitution occurring in response to the A&E target can be tested through the following hypotheses:

Hypothesis 4: The one-week measurement period for "star ratings" in March 2003 produced a temporary blip in performance that departed sharply both from the period before and the period after "sweeps week."

Hypothesis 5: Better performance in meeting the English A&E wait time target is associated with increased admission into inpatient wards.

We are unable to test the hypothesis about ambulances waiting outside A&E departments due to lack of available data.[16]

Pressures for effort substitution and gaming are more intense during incentive periods; for example, Harris and Bromiley (2007; see also Freeman and Gelber, 2007). find that the greater the proportion of a chief executive's compensation is in the form of stock options, the more likely the company is to restate its earnings later due to accounting irregularities. So our hypotheses also imply a significant coefficient for an interaction between the presence of an incentive period and effort substitution or gaming.

Finally, it should be noted that the Healthcare Commission, a government audit body, conducted regular audits on recordkeeping in each trust to look for outright data falsification in A&E departments. Though the audits reveal a non-trivial error rate in the paperwork of 11%, mistakes were overwhelmingly of a administrative nature. The audits

[16] There is data on wait times between calling for an ambulance and arrival of the ambulance to the place it was directed to come, but no data on wait times between pickup and entry of the patient into the A&E department. Furthermore, there are far fewer ambulance trusts in England than there are hospital trusts, and their boundaries cannot be crosswalked to hospital catchment boundaries.

revealed a very few instances of errors in time records but no evidence of systematic fraud (Department of Health, 2005).

DATA AND METHODS

Our data come from the U.K. Department of Health. In England, there are 155 local "hospital trusts" in our period, each of which manages the local hospital(s) and associated care centers, with funding almost entirely from the national government. The primary variable of interest is the percent of patients treated within four hours of arrival in each A&E department, recorded weekly in each trust in England. Our data begin in January 2003 and run through the beginning of September 2006.[17]

In addition to our primary series, we also have a number of additional measures collected at the hospital trust level that allow tests for effort substitution and gaming. These data are (with one exception) not collected weekly but only quarterly, from the fourth quarters of 2002 through 2005, unless indicated.

To test our hypotheses, we use the following data:

(1) For Hypothesis One (treatment quality), we first use death rates for patients presenting in the A&E department.[18] This is of course the most dramatic measure of treatment quality. In addition to having these data by quarter, we also have them by week from January 2003 through March 2006.

---

[17] We have data from some hospitals back to April 2002, but we only have complete data beginning in 2003. A change in the patient allocation procedures means that there are actually two such series: those for "Type 1" patients, and those for all patients. Patients who are not "Type 1" refer to patients seen in newly established alternative treatment facilities for low-grade ailments known as "walk-in centers." Beginning in October 2003, the Department of Health included these patients in the local hospital trust's attendance figures. In practice, the walk-in centers almost never make a patient wait more than four hours, and so the effect is to increase the denominator of the ratio which forms the dependent variable. The correlation between the two series is greater than 0.999, and so the choice of series does not substantively affect the results below. We thus use the "Type 1" patient series to make comparison across different periods easier.

[18] The data include deaths that occur in the same "spell," which may or may not include multiple visits to different parts of the hospital (so long as the patient originally presented in the A&E). Thus we may include deaths that occurred in inpatient wards, e.g. a patient presenting in the A&E department who then dies a few days later in an inpatient ward.

The data collected are for the number of deaths, not the death rate. Even though we conduct the following empirical analysis entirely using within-hospital variation, so that differences in size (and thus total number of deaths) across hospitals will not skew the analysis, hospitals may change in size throughout the period. In order that we not confuse an expanding hospital with one whose death rates are rising, we use the death rate, which we calculate as the ratio of deaths to the number of patients seen in the A&E during a given week.

Secondly for Hypothesis One, we use the number of patients returning to the A&E department within 30 days of a previous A&E visit ("return rates"), by quarter. This gives us another good measure of quality, since most return visits are in response to ineffective treatments.

(2) For Hypothesis Two (resource redirection from other elective activities), we use waiting times for elective orthopedic and trauma-related surgery. We have quarterly data from April 2002 through March 2006. Our data comprise snapshots of the waiting list, including the total number of patients and, within bins, how long each of those patients have been waiting. We use the average waiting time at each of our snapshots.

(3) For Hypothesis Three (redistribution of wait times), we use data for the distribution of wait times within one-hour bins up to four hours, which allows us to calculate the fraction of patients treated in under two hours as a fraction of all patients. Secondly, we may also use these same data to calculate an approximation to the mean wait time. To do so, we assume that patients in the zero to one hour bucket, on average, waited 30 minutes, and so on upwards; and that mean wait time for breaches of the four-hour target was five hours. In the basic specifications, we assume that the average

waiting time for those waiting more than four hours is five hours; our results are not qualitatively sensitive to changing this assumed value to four or six hours.

(4) For Hypothesis Four (blip during "sweeps week"), we examine the time series for mean wait-time performance beyond the last week of March 2003.

(5) For Hypothesis Five (increased admissions to inpatient), we use data on such admissions.[19]

Our method to test for effort substitution and gaming is to estimate regressions of the general form

$$y_{ht}=\alpha+\beta x_{ht}+\nu_{h}+\varphi_{t}+\varepsilon_{ht} \qquad (1)$$

where y_{ht} represents a non-targeted measure of performance, x_{ht} is the fraction of patients treated in under four hours, ν_{t} and φ_{h} are quarter- and hospital-specific fixed effects, for hospital h in quarter t. We also replace x_{ht} with $\Delta x_{ht}=x_{ht}-x_{h,t-1}$ in some specifications.

One concern with this specification is that moving from 96% to 98% might represent a greater improvement than one from 76% to 78%, since each breach of the 4-hour time limit is harder to remove than the last. We therefore also experiment with an alternatively scaled (logarithmically transformed) measure of performance x_{t}, where

$$x^*_{ht}=-\ln(1-x_{ht}).$$

[19] During the period covered by our data, many hospitals also introduced "clinical observation wards," which were inpatient facilities with lower standards than those in traditional hospital room (but higher than A&E departments), for A&E patients requiring tests that would take more than four hours. Patients admitted into these wards within four hours were considered to have met the four-hour target. However, such patients were coded as having been admitted to standard inpatient wards; thus, the extent that hospitals used these new wards to meet the standards is reflected in the inpatient admission figures.

Like actual performance, $x^*_{ht}>0$, and $x^*_{ht}=0$ when $x_{ht}=0$. This scales the independent variable to value proportional decreases in breaches equally; improving from 96% to 98% is now equivalent to improving from 76% to 88%.

In each of these specifications, the parameter β estimates the extent to which increases in targeted performance are associated with changes in the alternative performance measures. For all of our measures of effort substitution and gaming, an estimate of $\beta>0$ implies support for the hypothesis.[20] We cluster standard errors by trust.

Using each of these independent variables, we also run regressions using a specification with an interaction term

$$y_{ht}=\alpha+\beta x_{ht}+\gamma x_{ht}*I_{t}+\nu_{h}+\varphi_{t}+\varepsilon_{ht} \quad (2)$$

where I_{t} is a dummy variable equal to one in the pre-sweeps or cash incentive periods. The parameter γ estimates effort substitution or gaming by looking at the differential response rate to increases in the measured statistic between incentivized and non-incentivized periods. If $\gamma>0$, there is evidence of effort substitution or gaming. As before, we cluster standard errors by trust.

RESULTS

Descriptive Statistics

Table 1 presents descriptive statistics for the important variables in our analysis, Panel A shows the mean and standard deviation of these variables across our entire period, while Panel B displays the average values of these variables at the beginning and ending of our time period, to give a sense of the changes. (Since the quarters for which data are available vary somewhat across the different measures, we also present the

[20] So that lower measures of z are always better, we use the fraction of patients treated in more than two hours.

Table 1: Summary Statistics for Performance Measures

Panel A: Entire Sample

Variable	Mean	Std. Dev.	N
% of Hospitals Peforming > 98%	22.5%	-	2633
Hospital Deaths per A&E Admit	0.98%	0.37%	1996
% Patients on Return Visit	5.12%	3.53%	2002
Elective Surgery Waiting Times	3.33	0.94	2297
% Patients Treated < 2 Hours	56.7%	10.9%	1979
"Mean" Patient Wait Time	2.02	0.37	1979
% Patients Admitted to Hospital	18.53%	5.37%	2008

Panel B: by Period, Beginning and Ending

Period	Variable	Mean	Std. Dev.	N	Start/End Qtr.
	% of Hospitals Peforming > 98%	1.24%	-	155	Q3, 2002
	Hospital Deaths per A&E Admit	1.17%	0.43%	150	Q1, 2003
	% Patients on Return Visit	6.41%	4.47%	150	Q4, 2002
Beginning	Elective Surgery Waiting Times	4.52	0.86	150	Q3, 2002
	% Patients Treated < 2 Hours	47.4%	13.3%	144	Q4, 2002
	"Mean" Patient Wait Time	2.50	0.49	144	Q4, 2002
	% Patients Admitted to Hospital	18.7%	4.77%	150	Q4, 2002
	% of Hospitals Peforming > 98%	59.4%	-	155	Q3, 2007
	Hospital Deaths per A&E Admit	0.84%	0.32%	155	Q1, 2007
	% Patients on Return Visit	4.14%	2.98%	154	Q3, 2005
Ending	Elective Surgery Waiting Times	2.31	0.33	154	Q1, 2007
	% Patients Treated < 2 Hours	58.3%	10.6%	154	Q4, 2005
	"Mean" Patient Wait Time	1.88	0.27	154	Q4, 2005
	% Patients Admitted to Hospital	18.5%	5.22%	154	Q3, 2005

Panel A provides summary statistics all for all hospital-quarter observations. Panel B provides summary statistics for the first and last quarters of data for each variable in our sample. The final column denotes the timeperiod.

endpoints of the window in Panel B). Averaged across the entire sample, 22.5% of hospitals met the threshold of treated 98% of patients within four hours in each week, but this masks substantial improvement during our time period. Panel B shows that while only 1 or 2 hospitals (of 155) met the standard during the initial quarter of data, nearly 60% did so by the end of the sample.

Table 1 also presents descriptive statistics for our six measures of effort substitution and gaming, all of which improve during our sample period. Death rates decrease over time, from 1.24% of patients at the beginning of the sample to 0.84% by the end, as does the fraction of patients who come into the A&E on a return visit. Waiting times for orthopedic surgery decline dramatically, falling nearly in half over the three plus years of data. Mean wait time also decreases[21] from 2.50 hours to 1.88 hours, and the percent of patients treated within two hours increases as well.[22] Finally, the percent of patients admitted to inpatient wards decreases slightly.

Hypothesis Testing

Table 2 reports results from this set of 24 regressions; there are three different measures of quality of care (death rates by quarter, death rates by week, and return visit rates), four versions of the measured statistic (performance, change in performance, log performance, change in log performance), and two specifications (equations (1) and (2)). For each dependent variable, the first column presents the specification in Equation 1 and the second in Equation 2. Each vertical column displays β's for the regressions that share a dependent variable and a specification. For instance, in Column 1, β=0.481 when the

[21] For this summary statistic, we use the assumption that patients breaching the four-hour target were on average treated in five hours.

[22] The fraction of patients treated within one hour increases as well (see Friedman and Kelman, 2007).

Table 2: Testing for Effort Substitution: Death Rates and Return Visits

Explanatory Variables:	*Death Rates (by Qtr)* (1)	(2)	*Death Rates (by Week)* (3)	(4)	*% Patients on Return Visit* (5)	(6)
Performance	-1.190**	-1.138**	-0.654**	-0.636**	0.013	-0.018
	(0.239)	(0.253)	(0.146)	(0.150)	(0.169)	(0.037)
*Performance*Incentives*	-	-0.104	-	-0.032	-	0.034
		(0.190)		(0.150)		(0.036)
ΔPerformance	0.481**	0.345	-0.105	-0.078	-0.016	-0.046*
	(0.176)	(0.181)	(0.076)	(0.103)	(0.009)	(0.022)
*ΔPerformance*Incentives*	-	0.238	-	-0.052	-	0.037
		(0.297)		(0.155)		(0.023)
Log Performance	-0.046**	-0.030	-0.023*	-0.014	0.003	0.000
	(0.016)	(0.019)	(0.011)	(0.011)	(0.002)	(0.003)
*Log Performance *Incentives*	-	-0.037*	-	-0.018	-	0.004
		(0.017)		(0.010)		(0.003)
ΔLog Performance	0.003	0.015	0.000	0.002	-0.001	-0.004**
	(0.009)	(0.017)	(0.004)	(0.005)	(0.001)	(0.001)
*ΔLog Performance *Incentives*	-	-0.026	-	-0.005	-	0.005**
		(0.021)		(0.007)		(0.002)
Trust Effects?	yes	yes	yes	yes	yes	yes
Quarter Effects?	yes	yes	-	-	yes	yes
Week Effects?	no	no	yes	yes	no	no
N	1996 / 1996		26195 / 26195		2002 / 1841	

Statistical significance is denoted with the system: * 5%, ** 1%. Standard errors are clustered at the trust level. The number of observations records two sample sizes: that for the level regressions, and that for the difference regressions, respectively. The dependent variables, when percentages, are scaled to range from 0 to 100. Performance as an independent variable is scaled from 0 to 1.

dependent variable is the death rate, measured quarterly, the independent variable is the change in wait-time performance (Δx_{ht}), and the specification is that in Equation (1).

Hypothesis One states that improved wait times would be associated with lower quality of care. Instead, we find only two coefficients out of 24 that are statistically greater than 0, in the second panel of Column 1 and the fourth panel of Column 6. For 17 specifications there is no significant effect. And in five specifications, there is a significant effect, but it goes in the direction opposite to that predicted by the hypothesis. Both the better the wait-time performance, the lower the death rate, although the magnitudes of the improvements are somewhat small, though. An improvement of 10 percentage points in wait times across a quarter implies that death rates would fall by 0.1 percentage points, about one-third of a standard deviation. A similar sized effect is present in the weekly results. Furthermore, these regressions control fully for changes over time; thus, this regression shows that hospitals which improve most along the measured dimension also decrease their death rates fastest. There is some evidence that improvements in A&E performance lead to a temporary increase in death rates, based on the positive coefficient in the second panel of Column 1, but this is not born out in the weekly data.

For most of the specifications using return visits as the alternative measure of quality, there is also no statistically significant effect. One of the specifications confirms the hypothesis, but the magnitude is extremely small. And for two specifications where there is a significant effect, the effect goes in the opposite direction to that Hypothesis One predicts, with better wait-time performance associated with fewer return visits. These coefficients are also small in economic magnitude.

Thus, taking these two measures of quality of care, Hypothesis One is not confirmed. Instead, there is some evidence that what is occurring is the opposite to what is predicted by Hypothesis One: shorter waits are associated with better quality of care, not worse.

Hypothesis Two states that improved wait times would be associated with increased orthopedic wait times. Table 3 presents these results in columns 1 and 2. However, there are no specifications in which wait-time performance significantly affects orthopedic wait times. The magnitudes of the coefficients in either direction (orthopedic wait time increases or decreases) are also minute; for instance, in the first panel of column 1, the coefficient of -0.162 implies that a 10 percentage point improvement in measured performance decreases mean orthopedic wait at the end of the quarter by 0.016 hours, barely more than one-thousandth of a standard deviation. Hypothesis Two is not confirmed.

Hypothesis Three states that improved wait times would be associated with an increase in mean wait time. Table 3 presents these results in columns 3 through 6. However, all but one of the statistically significant coefficients in the regressions using mean wait time are negative, so that, as fewer patients wait more than four hours, mean wait time decreases.[23] Similarly, most of the significant coefficients in regressions using the sub-two hour fraction are negative. There are two significant coefficients that go in the other direction, in the second panel of columns 4 and 6, both involving interaction terms that compare incentive periods with periods where there were no special performance incentives. Relative to a similar move in periods without incentives,

[23] This result is robust to making the extremely conservative assumption that breached patients wait only four hours, on average.

Table 3: Testing for Effort Substitution: Waiting Times and Hospital Admissions

Explanatory Variables:	*Dependent Variable:* *Surgery Wait Times* (1)	(2)	*"Mean" Wait Time* (3)	(4)	*% Waits > 2 Hours* (5)	(6)	*Hospital Admits* (7)	(8)
Performance	-0.162	-0.472	-3.207**	-3.186**	-0.662**	-0.659**	-0.030	-0.046
	(0.331)	(0.496)	(0.180)	(0.236)	(0.067)	(0.088)	(0.021)	(0.032)
*Performance*Incentives*	-	0.351	-	-0.028	-	-0.004	-	0.018
		(0.389)		(0.202)		(0.071)		(0.026)
ΔPerformance	-0.024	0.753	-0.527**	-1.369**	-0.105	-0.345**	0.003	-0.018
	(0.257)	(0.419)	(0.198)	(0.196)	(0.061)	(0.070)	(0.014)	(0.034)
*ΔPerformance*Incentives*	-	-0.937	-	1.053**	-	0.275**	-	0.025
		(0.491)		(0.247)		(0.085)		(0.039)
Log Performance	0.000	0.024	-0.221**	-0.184**	-0.050**	-0.044**	-0.002	-0.001
	(0.028)	(0.032)	(0.018)	(0.020)	(0.006)	(0.007)	(0.003)	(0.002)
*Log Performance *Incentives*	-	-0.042	-	-0.067**	-	-0.012*	-	-0.001
		(0.039)		(0.018)		(0.006)		(0.002)
ΔLog Performance	-0.006	0.031	-0.107**	-0.102**	-0.028**	-0.029**	-0.002	0.000
	(0.017)	(0.031)	(0.008)	(0.012)	(0.003)	(0.004)	(0.001)	(0.002)
*ΔLog Performance *Incentives*	-	-0.057	-	-0.010	-	0.002	-	-0.002
		(0.041)		(0.017)		(0.006)		(0.003)
Trust Effects?	yes	yes	yes	yes	yes	yes	yes	yes
Quarter Effects?	yes	yes	yes	yes	yes	yes	yes	yes
N	2297 / 2147		1979 /1979		1979 /1979		2008 / 1847	

Statistical significance is denoted with the system: * 5%, ** 1%. Standard errors are clustered at the trust level. The number of observations records two sample sizes: that for the level regressions, and that for the difference regressions, respectively. The dependent variables, when percentages, are scaled to range from 0 to 100. Performance as an independent variable is scaled from 0 to 1.

increasing performance by one percentage point increases the fraction of patients waiting more that two hours by 0.275 percentage points. This effect is less than 2.5% of one standard deviation of this variable, though. Similarly, a one percentage point increase in wait-time performance increases mean wait time by 0.01 hours more in incentive than in other periods. Thus, Hypothesis Three is not confirmed; instead, there is again evidence, here quite strong, that what is occurring is the opposite: shorter waits are associated with lower mean wait times and a higher fraction of patients treated in under two hours.

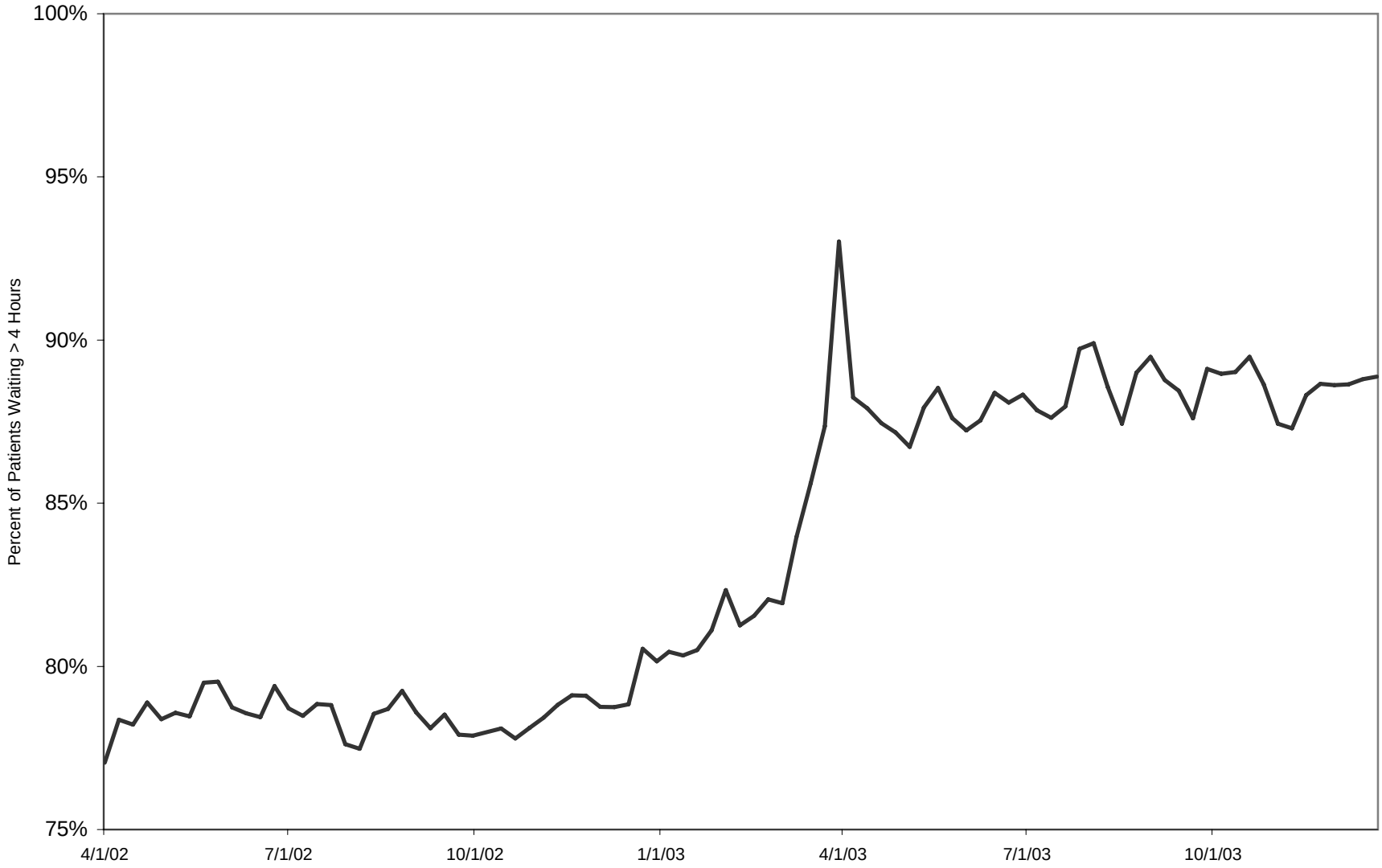
The fact that the results are the opposite of those predicted by Hypothesis Three helps explain the fact presented in Table 1 that as attainment of the four-hour threshold improved, average wait times improved dramatically: as noted, we estimate that average A&E wait times declined by one-third between December 2002 and December 2005 (from 2.81 to 1.88 hours), a dramatic performance improvement. Put another way, the 10th percentile of wait-time performance in the final period lies above the 90th percentile of performance in the first.

Hypothesis Four states that the performance improvement associated with the March 2003 star ratings "sweeps week" would be a temporary blip. Figure 2 displays the time series for mean wait-time performance beyond the last week of March 2003. As can be seen from this continuation of the time series, much, though not all, of the

[FIGURE 2 GOES ABOUT HERE]

performance improvement occurring during that week persisted thereafter. This time series that received no media attention when, several months later, the performance spike during that was noted in the media.

Figure 2: Emergency Room Waiting Times



During sweeps week, compliance rates spiked from an average of 84.0% for the three previous weeks to 93.0%. Though performance fell by nearly five points in the following week, it remained nearly constant over the next nine months at a level far above pre-sweeps performance. "Sweeps week" produced a permanent performance improvement. A longer time series clouds considerably the gaming story the British media told at the time.[24] Hypothesis Four is not confirmed.[25]

Hypothesis Five states that improved wait times would be associated with an increase in admissions into inpatient wards. Results form regessions testing this hypothesis are in columns 7 and 8 of Table 3. Once again, though, there are no specifications in which wait-time performance significantly affects admission into inpatient wards, and all coefficients are of minute economic magnitude. Hypothesis Five is not confirmed.

DISCUSSION

In the case of the English A&E wait-time target, none of the hypotheses predicting effort substitution or gaming in connection with attaining this target has been confirmed. To the extent there are significant results (regarding death rates and mean/under two hour wait times), they go in directions opposite to those predicted by effort substitution and gaming stories.

[24] Something else that contemporaneous press accounts did not note is that actual performance improvement for in A&E departments started to occur very shortly after the announcement of the government's new attention to this target, several months before the measurement week. We return to this question in the next section of this paper.

[25] Friedman and Kelman (2007) present a number of econometric tests that confirm the general story the simple time-series figure shows.

Given the theoretical arguments for why we might expect dysfunctional responses to performance measures, this section – in the spirit of the saw that an economist is someone who, seeing that something works in practice, asks why it would work in theory – presents a theoretical discussion about why, or stated better, under what circumstances, dysfunctional responses to performance measurement do not arise. We discuss (1) complementarity across performance dimensions and (2) ways that dysfunctional responses become self-limiting, (3) management behaviors to limit dysfunctional responses.

Complementarity

Effort substitution is a problem if the two dimensions of effort substitute for each other. It is not a problem if they complement each other, i.e. if improvement in the measured dimension at the same time improves performance in the unmeasured dimension. Holmstrom and Milgrom recognize this in their 1991 paper. Discussing the two tasks of teaching basic and higher-thinking skills (the former assumed to be measurable, the latter not), they note (pp. 32-33) that when these activities "are complementary in the agent-teacher's private cost function, the desirability of rewarding achievement in teaching basic skills is enhanced," as opposed to a situation where "the two dimensions of teaching are substitutes." There is evidence, for example, that creation of a stable and task-directed classroom environment, which aids teaching topics for standardized tests, promotes classroom learning in general (Schmidt et al, 1999; Rowan, Correnti & Miller, 2002). Jacob (2005) found that, while scores in reading and math, the tested subjects, rose in Chicago in response to introduction of school testing in the 1990's,

scores in science and social studies rose as well (the increases were statistically significant), though considerably more slowly than reading and math scores.

Some of the failure to observe dysfunctional effects in the A&E wait time target would appear to involve situations where there are complementaries. We have argued elsewhere (Friedman and Kelman, 2007) that the reason wait-time performance organizational and procedural improvements made to improve performance during that week carried over, once having been made, to post-measured week performance. (Similarly, performance began to improve immediately following the Department of Health's January 2003 announcement of the March star ratings measurement week, due, we argue, to a similar complementarity whereby hospitals were "practicing" for performance during that week.) Thus, improving the measured week performance was complementary to improving performance both after and before the measurement. In the case of the positive relationships between speed and quality of care (as reflected in A&E death and return rates), one may note that speedy treatment ceteris paribis is likely to improve care outcomes, because many conditions leading a patient to present for emergency care (though not all) will get worse if not treated promptly – another example of complementarity.

As for the positive relationship between improvement on the four-hour target and increase in the percentage of short waits, in connection with the focus on meeting the four-hour target the Department of Health publicized, organized training, and consulted with poorly performing hospitals on various organizational and procedural changes recommended as ways for A&E departments to meet the target (e.g. Alberti, 2004). Particularly prominent was an idea called "see and treat," which proposed redesign of

traditional triage procedures so as to treat minor injuries more quickly. Under the previous triage system, a low-priority patient (for instance, someone with a minor wound requiring stitches) might wait for many hours until all more serious patients were treated; under "see and treat," A&E departments were urged to assign nurse-practitioners to deal with such injuries immediately. Another business-process change, called "wait for a bed," involved another source of long A&E waits, which was that a patient who was to be admitted to an inpatient ward needed to wait a long time in the less-attractive conditions of the A&E department while waiting for an (unavailable) inpatient bed. Here, the effort was to achieve better scheduling of inpatient operations and releases, coordinated with times when there is a large demand for inpatient beds from patients presenting through A&E (for example, on Saturday nights), so as to maximize the probability inpatient beds would be available when A&E needed them.

What should be noted about both these business-process changes that were recommended for improving attainment of the four-hour target is that they also improved short-wait performance, again creating a complementarity. "See and treat" dealt with a cause of long waits by getting minor-injury patients out of A&E very quickly, reducing average wait times as well as compliance with the four-hour threshold. "Wait for a bed" involved improving overall bed availability, not micro-efforts to make a specific bed available for a specific patient who was about to breach the four-hour target, so its benefits applied to patients whenever they were ready for inpatient admission, whether after four minutes or four hours.

We suspect that the number of situations where activities are substitutes more than complements is greater than the number where the opposite obtains. We also

suspect that, the closer the two goals are, the more likely the goals are to be complements rather than substitutes, because the more similar the technologies involved in producing the goals are likely to be. But these are of course empirical questions, and situations need to be examined in each case to see whether substitution or complementarity prevails.

Self-Limitation

Self-limitation can occur when a change produces negative feedback, which Jervis (1997: 125) defines as a situation where a change "treiggers forces that counteract the initial change." If dysfunctional responses generate negative feedback, this can limit the impact of those effects over time.

Dysfunctional responses may become self-limiting if those responses create problems for other subunits of the organization and lead those other subunits to object. Airline pilots are paid based on scheduled flight times and are limited in the amount of scheduled hours they may fly a month; gaming the on-time performance measure by padding scheduled flight times increases salaries pilots get paid and create a risk they may be excluded from flying at the end of a month (Gormley and Weimer, 1999: 149). In the A&E case, unnecessary admissions to inpatient wards from A&E create a burden for managers of inpatient wards, both in terms of capacity and of costs that are being shifted to them, and the self-limitation this creates is the likely explanation for the failure to find negative impacts of improving A&E wait-time performance on inpatient admissions.

Managerial Behaviors

Managers seeking to limit dysfunctional consequences of performance measures have several tools available. These include: (1) adding measures, (2) adapting measures, (3) cultivating public service motivation among employees.

The most obvious remedy for effort substitution across goals is to add an additional target to counteract substitution. The most obvious explanation for the failure to see effort substitution from orthopedic surgery into A&E departments is that there existed a target for reducing elective surgery wait times (which empirically turn out to be heavily driven by orthopedic wait times) at the same time as the A&E target was present. If there is worry about ambulances be kept waiting outside the A&E until they are ready to take patients, one could add an ambulance performance measure for average time from when the ambulance picks up the patient to when the patient is registered at the A&E.

This is not a perfect solution. The whole point of the tradition of economic theorizing about effort substitution beginning with Holmstrom and Milgrom is that sometimes it is difficult to develop good performance measures for all important dimensions of performance. Furthermore, there is a tradeoff between the focusing benefits that performance measurement seeks and the proliferation of measures this approach may create. Nonetheless, when available, this is an easy solution to effort substitution problems.[26]

A second managerial behavior is to adapt measures over time to reflect organizational learning about gaming responses. For example, Texas changed its eligibility rules in 1999 to require special education students to take the standardized tests

[26] To counteract de-focusing effects, one could make these additional measures into minimum standards, where the intention is not that more is better, but simply that the second measure be a constraint on achievement of the primary measure.

(Cullen and Reback, 2006).[27] Skeptics will see this as an example of cascading regulation (Zeckhauser, 1979), ultimately futile or self-defeating; Pollitt (1990) notes that Soviet planners engaged in a constant race against gaming by adapting and complexifying performance measures. However, the more sympathetic observer will see this as illustrating the evolution of rules over time in response to learning new information (March, Schulz, and Zhou, 2000).

A third managerial behavior can be to harness the existing public service motivation/intrinsic motivation (Deci et al, 1999; Perry and Wise, 1990; Crewson l997; Brehm and Gates, l999) and/or the professional values of one's employees, against gaming responses. A manager seeking to harness public service motivation against gaming would point out to employees that gaming does nothing to improve real performance and thus runs counter to the service or mission goals of the organization.[28] Note the different results in for Texas (noted earlier), compared with those in Boyne and Chen (2007) for England, regarding gaming standardized school tests by excluding pupils from the pool of those taking the test – whereas the Texas papers found gaming, Boyne and Chen found no relationship between the percentage of students excluded from the test and performance improvement. The difference in these results might be due to different levels, or different levels of mobilization by managers, of public service motivation in the different organizations. If there are otherwise incentives to game, mobilization of public-service motivation is unlikely completely to eliminate gaming responses, but it may reduce its magnitude.

---

[27] The authors (p. 6) note laconically that their "analysis is based on data from the years leading up to these reforms when gaming is likely to have been more prevalent."

[28] Public-service motivation and/or professional values might also be mobilized against dysfunctional effort substitution, such as providing poor quality of care so as to move patients through the system more rapidly.

We conclude with a theoretical observation regarding dysfunctional responses and performance improvement. Simply to note that a performance measurement regime produces some level of dysfunctional response does not by itself imply that such a regime fails on balance to improve organizational performance. The appropriate comparison is not between an organization's performance level assuming performance measurement with no dysfunctional responses and the level with some dysfunctions; the former will usually be greater than the latter. The appropriate comparison is between an organization's performance level assuming performance measurement assuming the dysfunctional responses and the counterfactual performance level with no measurement . If the former is greater than the latter, it is better to manage an organization using imperfect performance measures than using none at all.[29]

Holmstrom and Milgrom's argument is deductive -- their paper is trying to explain the presence of non-incentivized pay regimes, and the argument in effect takes the form that measurable performance dimensions must be less important than unmeasurable ones, or else the existence of non-incentivized pay wouldn't be rational. However, this is of course an empirical question (along, in government where the different performance dimensions are typically not monetized, with a values question about how important different performance dimensions are). Say the increased value of performance produced by the change in effort in response to incentivizing the measured dimension is much more important than the value of the performance decrement along

[29] Pollitt (1990: 172) makes this point, although with an emphasis on decisions rather than performance levels: "Of course it is easy to criticise the incompleteness of [performance] indicators, and to stress the perverse incentives which may be created if too much weight is put upon them. But to be convincing, the analysis must surely always be comparative: how would decision making with such partial and tentative indicators of outcomes compare with our current decision-making without them? What are the perverse incentives which are built into our current, possibly highly impressionstic and unreliable decision criteria?"

unmeasured dimensions. Then it makes sense to incentivize that dimension and take the effort substitution "hit."[30] The same goes for gaming. Blau (1955: 36) is positive on the whole towards performance measurement because, despite the dysfunctional reactions he notes, the percentage of job openings filled in the state employment agency he studied increased from 55% before introduction of the measures to 67% two months after they were introduced. In the Schweitzer, Ordonez, & Douma experiment discussed earlier, the percentage of experimental subjects who actually met the goal increased from 12% under the "do your best" condition to 24% under the reward goal condition; one might regard the finding that somewhat less than one in eight subjects who didn't meet the goal claimed to have done so (meaning that more than seven of eight reported truthfully that they hadn't met it) as suggesting relatively low levels of cheating, especially since subjects believed their representations could not be audited.

Thus, even when dysfunctional responses to performance measurement in government occur, the appropriate policy response may no more be to eliminate a performance measurement regime than would the appropriate response to Enron be to eliminate using profit as a performance measure to improve the performance of firms.

---

[30] Although there is clearly no assumption to this effect in the economics literature on effort substitution – which generally assumes that incentivizing the measured performance dimension will increase net effort compared to a non-incentivized world – the public management literature addressing effort substitution often assumes in effect there is a "lump of effort" to be allocated across, say, two activities, such that any observed increase in effort on the measured activity is necessarily matched by an equal reduction of effort on the non-measured activity.

In the case of the Holmstrom and Milgrom theoretical argument, one might ask what reason there is to believe that the unmeasurability of a performance dimension should be positively correlated to its importance.

REFERENCES

Alberti, George. 2004. *Transforming Emergency Care in England*. London: U.K. Department of Health.

Argote, Linda. 1999. *Organizational learning: Creating, retaining, and transferring Knowledge*. New York: Springer.

Baker, G. P. 1992. Incentive contracts and performance measurement. *Journal of Political Economy* 100(3): 598–614.

Berliner, Joseph S. 1956. A problem in soviet business management. *Administrative Science Quarterly* 1: 86–101.

Bevan, Gwyn, and Christopher Hood. 2006. What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration* 84(3): 517–538.

Blau, Peter M. 1955. *The dynamics of bureaucracy*. Chicago: University of Chicago Press.

Bohte, John, and Kenneth J. Meier. 2000. Goal displacement: Assessing the motivation for organizational cheating. *Public Administration Review* 60(2): 173–182.

Boyne, George A. and Alex A. Chen. 2007. Performance targets and public service improvement. *Journal of Public Administration Research and Theory* 17(3): 455–477.

Brehm, John O., and Scott Gates. 1999. *Working, shirking, and sabotage: Bureaucratic response to a democratic public*. Ann Arbor, MI: University of Michigan Press.

Carvel, John. 2003. Hospitals faked wait times test. *The Guardian* (London). May 13, 2003. p. 10.

Carr-Brown, Jonathan, and Dominic Tonner. 2003. Hospitals fake casualty waiting times for tests. *Sunday Times* (London). May 11, 2003. p 4.

Cartern, Neil, Rudolf Klein, and Patricia Day. 1992. *How organisations measure success: The use of performance indicators in government.* London and New York: Routledge.

Courty, Pascal, and Gerald Marschke. 2004. An empirical investigation of gaming responses to explicit performance incentives. *Journal of Labor Economics* 22(1): 23–56.

Crewson, Philip E. 1997. Public-service motivation: Building empirical evidence of incidence and effect. *Journal of Public Administration Research and Theory* 4, 499–518.

Cullen, Julie Berry and Randall Reback. 2006. Tinkering toward accolades: School gaming under a performance accountability system. In *Improving School Accountability:*

Check-ups or Choice, eds., Timothy J. Gronberg and Dennis W. Jansen. Amsterdam: Elsevier, 1–34

De Bruijn, Hans. 2007. *Managing Performance in the Public Sector*, 2nd ed. London and New York: Routledge.

Dechow, P. M., and D. J. Skinner. 2000. Earnings management: Reconciling the views of accounting academics, practitioners, and regulators. *Accounting Horizons* 14(2): 235–250

Deci, Edward L. et al. 1999. A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin* 125: 627-68.

Department of Health. 2005. *A&E Update, 10/12.* Section 5.1.1.

Department of Health. 2000. *The NHS Plan: A Plan for Investment, A Plan for Reform.* Available at http://www.nhsia.nhs.uk/nhsplan/nhsplan.htm

Figlio, David N. 2005. *Accountability, ability and disability: Gaming the system*. National Bureau of Economic Research Working Paper 9307. Cambridge, MA.

Figlio, David N. 2006. Testing, crime and punishment. *Journal of Public Economics* 90(4-5): 837–851.

Freeman, Richard B. and Alex M. Gelber. 2006. Optimal Inequality/Optimal Incentives: Evidence from a Tournament. NBER Working Paper No. 12588. Cambridge, MA: National Bureau of Economic Research.

Friedman, John N. and Steve Kelman. 2007. *Effort as Investment: Analyzing the Response to Incentives.* Kennedy School of Government Faculty Research Working Paper Series, RWP07-024, May 2007.

Gibbons, Robert. 1998. Incentives in organizations. *Journal of Economic Perspectives* 12(4): 115–132.

Gormley, William T. Jr., and David L. Weimer. 1999. *Organizational report cards*. Cambridge, MA: Harvard University Press.

Grizzle, Gloria A. 2002. Performance measurement and dysfunction. *Public Performance and Management Review* 25(4): 363–369.

Harris, Jared and Philip Bromiley. 2007. Incentives to cheat: The influence of executive compensation and firm performance on financial misrepresentation. *Organization Science* 18(3) 350–367.

Hatry, Harry P. 1999. *Performance measurement: Getting results*. Washington DC: Urban Institute Press.

Healy, Paul M., and James M. Wahlen. 1999. A review of the earnings management literature and its implications for standard setting. *Accounting Horizons* 13(4): 365–383.

Healy, Paul M. 1985. The effect of bonus schemes on accounting decisions. *Journal of Accounting and Economics* 7: 85–107.

Heckman, James J., Carolyn Heinrich, and Jeffrey Smith. 2002. The performance of performance standards. *The Journal of Human Resources* 37(4): 778–811.

Hedlund, Gunnar. 1994. A model of knowledge management and the N-form corporation. *Strategic Management Journal* 15 (Special Issue: Summer): 73–90.

Heinrich, Carolyn J. 2003. Measuring public sector performance and effectiveness. In *Handbook of Public Administration*, eds. B. Guy Peters and Jon Pierre. Thousand Oaks, CA: Sage.pp.25–37

Heinrich, Carolyn J. 1999. Do government bureaucrats make effective use of performance management information? *Journal of Public Administration Research and Theory* 9(3): 363–393

Hirschman, Albert O. 1991. *The Rhetoric of Reaction*. Cambridge, MA: The Belknap Press of Harvard University Press.

Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7 (Special Issue: Papers from the Conference on the New Science of Organization): 24–52.

Huber, George P. 1991. Organizational learning: The contributing processes and the literatures. *Organization Science* 2 (No.1, Special Issue: Organizational Learning: Papers in Honor of [and by] James G. March): 88–115.

Ilgen, Daniel R., Cynthia D. Fisher, and M. Susan Taylor. 1979. Consequences of individual feeback on behavior in organizations. *Journal of Applied Psychology* 64(4): 349–371.

Jacob, Brian A. 2005. Accountability, incentives and behavior: The Impact of high-stakes testing in chicago public schools. *Journal of Public Economics* 89: 761–796.

Jacob, Brian A., and Steven D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118 (3): 843–877.

Jensen, Michael C. 2003. Paying people to lie: The truth about the budgeting process. *European Financial Management* 9(3): 379–406.

Jensen, Michael C., and William H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Jervis, Robert. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, NJ: Princeton University Press.

John P. Jumper. 2000. Presented in a briefing called "What I Believe," to commanders of U.S. Air Force Air Combat Command. Washington: powerpoint.

Kelman, Steve. 2006. Improving service delivery performance in the United Kingdom: Organization theory perspectives on central intervention strategies. *Journal of Comparative Policy Analysis* 8(4): 393–419.

Kerr, Steven. 1975. On the folly of rewarding A, While hoping for B. *Academy of Management Journal* 18: 769–783.

Kravchuk, Robert S., and Ronald W. Schack. 1996. Designing effective performance-measurement systems under the Government Performance and Result Act of 1993. *Public Administration Review* 56(4): 348–358.

Locke, Edwin A., and Gary P. Latham. 1990a. Work motivation: The high performance cycle. In *Work Motivation,* eds. Uwe Kleinbeck et al. Hillsdale, NJ: Lawrence Erlbaum, 3–25.

Locke, Edwin A., and Gary P. Latham. 1990b. *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.

Locke, Edwin A., and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motiviation: A 35-year odyssey. *American Psychologist* 57(9): 705–717.

March, James G., Martin Shultz, and Zueguang Zhou. 2000. *The Dynamics of Rules*. Stanford, CA: Stanford University Press.

McNichols, Maureen F. 2000. Research design issues in earnings management studies. *Journal of Accounting and Public Policy* 19: 313–345.

McNichols, Maureen, and G. Peter Wilson. 1988. Evidence of earnings management from the provision for bad debts. *Journal of Accounting Research* 26 (Supplement): 1–31.

Meikle, James. 2003. Waiting times in A & E 'fiddled'. *The Guardian* (London). March 29, 2003, p. 15.

Merton, Robert. 1936. The unanticipated consequences of purposive social action. *American Sociological Review* 1 (December): 895–904

Metzenbaum, Shelley. 2003. *Strategies for using state information: Measuring and improving program performance*. Washington DC: IBM Center for the Business of Government.

Meyer, Marshall W. and Vipin Gupta. 1994. The performance paradox, *Research in Organizational Behavior* 16: 309–369.

Perry, James L., and Lois R. Wise. 1990. The motivational bases of public service. *Public Administration Review* 50: 367–73.

Talbot, Colin. 2005. Performance management. In *The Oxford Handbook of Public Management*, eds. Ewan Ferlie, Laurence E. Lynn Jr. and Christopher Pollitt. Oxford: Oxford University Press, 491–517.

Radin, Beryl A. 2006. *Challenging the performance movement*. Washington DC: Georgetown University Press.

Rainey, Hal. 1993. Toward a Theory of Goal Ambiguity in Public Organizations. In James Perry, ed. *Research in Public Administration*, Vol. 2. Greenwich, CT: JAI Press, pp. 121–166.

Ridgeway, V. F. 1956. Dysfunctional consequences of performance measurements. *Administrative Science Quarterly* 1(2): 240–247.

Rowan, Brian, Richard Correnti & Robert J. Miller. 2002. What large-scale, survey research tells us about teacher effects on student achievement: Insights from the prospects study of elementary schools. *Teachers College Record* 104 (8): 1525-1567.

Schmidt, William H., Curtis C. McKnight, Leland S. Cogan, Pamela M. Jakwerth, Richard T. Houang with the collaboration of David E. Wiley, Richard G. Wolfe, Leonard J. Bianchi, Gilbert A. Valverde, Senta A. Raizen, and Christine E. DeMars. 1999. *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education*. Dordrecht, Boston: Kluwer Academic Publishers.

Schweitzer, Maurice E., Lisa Ordonez, and Bambi Douma. 2004. Goal setting as a motivator of unethical behavior. *Academy of Management Journal* 47(3): 422–432.

Smith, Peter. 1995. On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18(2 & 3): 277–310.

Timmins, Nicholas. 2003 Hospitals make frantic efforts to hit A & E targets. *Financial Times of London*. March 29, 2003. p. 8.

Van Thiel, Sandra, and Frans L. Leeuw. 2002. The performance paradox in the public sector. *Public Performance and Management Review* 25(3): 267–281.

Wilson, James Q. 1989. *Bureaucracy: What government agencies do and why they do it*. New York: Basic Books.

Winter, Søren C. 2005. Effects of Casework: The Relation between Implementation and Social Effects in Danish Integration Policy. Unpublished paper presented at the 2005 Research Conference of the Association for Public Policy and Management. Washington, DC: November 3-5, 2005.

Zeckhauser, Richard J. 1979. Using the Wrong Tool: The Pursuit of Redistribution through Regulation. Unpublished Paper prepared for the U.S. Chamber of Commerce, Council on Trends and Perspective. Cambridge, MA: Kennedy School of Government.