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

ABSTRACT

The literature on the use of performance measurement in government has focused much attention on hypothesized unintended dysfunctional consequences that 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 the establishment in the English 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 hospital trusts 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, PERFORMANCE TARGETS, 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. 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

Acknowledgements Anonymized

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 previously thrown dart 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 – in fact, it has often focused on dysfunctional responses as much as with functional. These worries evoke the spirit, and often the letter, of Merton's (1936) idea of the "unintended consequences of purposive social action." One paper (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 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.)

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. 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.

In this paper we explore the question of dysfunctional responses to a performance target during the Blair government in the United Kingdom for wait times in English[1] hospital "Accident and Emergency" (A&E) departments -- equivalent to emergency rooms in the United States – run by the governmental National Health Service. As we shall see, wait-time performance in English A&E departments improved dramatically during the years between 2003 and 2006 in which government focused on this target. In the context of this performance improvement, the paper has four aims. First, we discuss theoretically – using literature from public management, economics, organization theory, and accounting – why one might expect dysfunctional responses to 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

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

on econometric analysis of data from all English hospitals during the period 2003-06, on presence of the predicted dysfunctional effects. We find no evidence of these dysfunctional responses. 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 wait-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 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 possible presence of distortionary effects.

BACKGROUND

England's National Health Service (NHS) was established in 1948 as the nation's primary healthcare system.[2] 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[3] 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

[2] This section closely follows ANONYMIZED (2007).
[3] Or for admission into an inpatient ward.

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 established in the U.K.[4] The first such ratings for hospitals were prepared in 2001. 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 created to work on targets, released a plan for meeting the A&E target. This document announced monetary incentives for hospitals that met the target, along with ongoing monitoring of weekly performance data 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. [5]

---

[4] 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.

[5] 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.

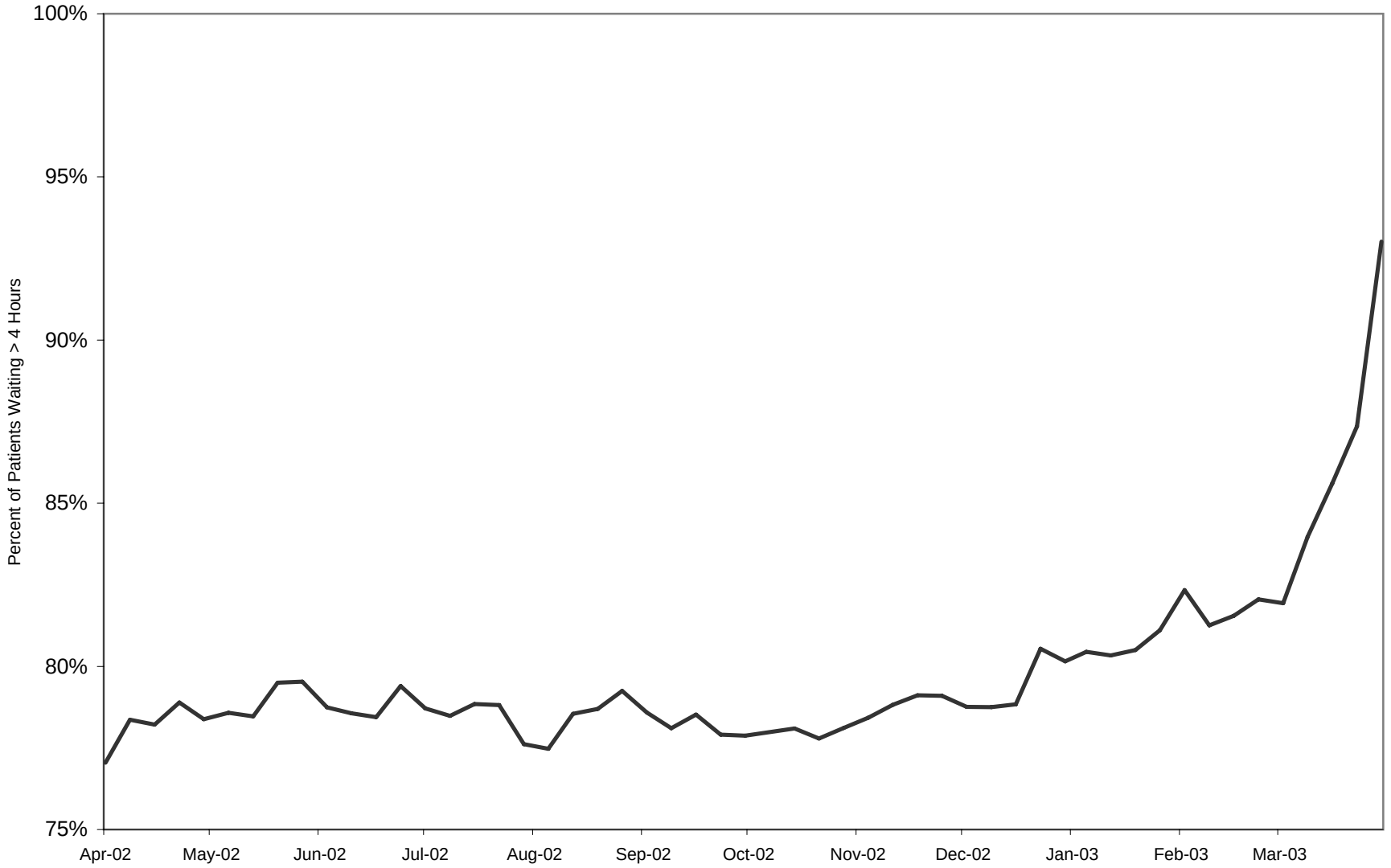
One gets a flavor for concerns about dysfunctional responses to the Blair government's performance targets by examining the first major effort undertaken to encourage attainment of the A&E target, inclusion of A&E wait-time performance in 2003 hospital star ratings. As noted earlier, performance for the purpose of the star rating was measured during one week, announced well in advance; elsewhere (ANONYMIZED, 2007) we have referred to this measurement week as a "sweeps week" for the hospitals.[6] During that week, A&E wait-time performance spiked up dramatically from that during previous weeks and months. This is seen in Figure 1, which shows average wait-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," it was 93%.

[FIGURE 1 GOES ABOUT HERE]

Contemporaneous media coverage was quick to notice this 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 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

[6] 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



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.

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 (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).[7]

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.[8] 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 volume of output at the expense of quality."[9] Since Holmstrom and Milgrom, in economics this problem has generally been named "multitasking" or "effort substitution." 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

[7] 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."

[8] 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.

[9] 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."

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 reduction in instruction in social studies or foreign languages (typically not subject to 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 often pursue, and arguably because in the public sector more goals are difficult to quantify, or involve improvements whose effects appear only after many years (Rainey, 1993; Smith, 1995).

The more peripheral the aspect of performance a measure captures, the more 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, where the measure used is only distantly related to the underlying goal being sought.

For A&E departments, the most dramatic example of effort substitution would be to improve 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 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[10]; 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

[10] 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 care quality, especially when patients have life-threatening conditions requiring rapid attention.[11]

The presence of effort substitution 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."[12] 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.

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

[12] 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 tests 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 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). Alternatively, product shipment dates may be manipulated to fit into a certain month or quarter (Jensen, 2003). 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. Courty and Marschke 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. Finally, if gaming creates an incorrect impression of acceptable performance, the performance-promoting pressures that performance measurement produces will be short-circuited.

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.[13]

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 A&E (say with a heart attack), they clearly cannot be fully treated within four hours, so the target specifies they must be admitted to inpatient wards, where further treatment will take place, within four hours. It has been suggested (Bevan and Hood, 2006; Mayhew and Smith, 2008) 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

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

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 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.[14]

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

[14] 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.

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, has conducted audits on recordkeeping in each trust to look for data falsification in A&E departments. Though the audits reveal a non-trivial error rate in the paperwork of 11%, mistakes were overwhelmingly of an administrative nature. The audits revealed a very few instances of errors in time records and 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. In addition to our primary series, we have a number of additional measures collected at the hospital 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.[15] 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.

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 wait times for elective orthopedic and trauma-related surgery. This kind of care was chosen for two reasons. First, looking at elective surgery, resource redirection would be easier, since it would not involve removing resources from acute needs. Second, A&E departments receive a large number of orthopedic cases, arising from accidents, especially traffic accidents, and from domestic violence or other fights, and thus efforts to

[15] 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.

reduce wait times might especially be expected to involve transferring orthopedic care resources into A&E, making elective orthopedic surgery a most-likely case scenario for resource redirection[16] 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 wait 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 wait 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.[17]

[16] Trauma and orthopaedic surgery makes up 26% of all elective surgeries. We also estimate the specifications in Tables 3 and 4 using all elective surgery, and the results are qualitatively unchanged.

[17] 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.

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.[18]

Equation (1) relates the level of performance on the measured task to the level of performance on alternative tasks. This specification is appropriate for the standard model of effort substitution, in which the two tasks are substitutes in the agent's production function. But effort might not only relate to the level of performance but also to the changes in performance. For instance, such changes often require redesigning procedures or organizations, in which case effort functions more like investment in a traditional model (see ANONYMIZED, 2008 for more discussion and evidence). In this case, effort substitution would appear as a reduction in the level of performance on alternative tasks *at the time of an increase in measured performance*. To test for this possibility, we also replace x_{ht} with $\Delta x_{ht}=x_{ht}-x_{h,t-1}$ in some specifications.

Another 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

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

---

[18] As will be noted below in a different context, hospital budgets were increasing during this period. Any effects of these on our regressions are controlled for by the dummy variables for each time period, and any effects of variations in budget increases is controlled for by the hospital-level fixed effects.

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.[19] 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 endpoints of the window in Panel B).

Performance against the four-hour target improved dramatically during this period.[20] Averaged across the entire sample, 22.5% of hospitals met the threshold of

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

[20] It should be noted that both budgets and staffing for the National Health Service, including A&E departments, increased significantly during this period (Healthcare Commision, 2005a), and we do not claim that all the overall improvement reflected in performance was due to attention to the target; rather, some may reflect increased budgets. Several points should be noted, however. First, funding for the National Health Service began increasing dramatically starting in 1997, when the Labour government came to power, without having much impact on A&E wait times (or a number of other health system targets) before attention to the target starting in 2003. Second, looking at increases in A&E staff specifically (data exist for only two periods, 2000 and 2004), the Healthcare Commission (2005a) found no bivariate relationship between staff increases and wait-time improvements during those years. More broadly, there are certainly many examples – such as following the dramatic increase in per capita education spending in the decades following the 1960's – where large spending increases in government have not been matched by any noticeable performance improvements (Hanushek, 1996). We do not pursue this issue further

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.

treating 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

---

because explaining the performance improvement is not the topic of the empirical analysis this paper. As noted above (see footnote 18), our use of time-specific dummies and hospital-level fixed effects controls for impacts of budget increases on the phenomena we investigate here.

Additionally, Bevan and Hood (2006) note that surveys by government audit organizations of patient perceptions of A&E waiting times in 2003 and 2004 showed that 31% and 23%, respectively, of patients perceived having waited more than four hours (Commission on Healthcare Improvement, 2004; Healthcare Commission, 2005b), far higher than the figures reported here. If these perceptions reflect the reality of patient treatment, then the improvement numbers reported here are significantly exaggerated; results regarding effort substitution and gaming would be biased, however, only if the difference between our statistics and perceptions were caused by fraudulent reporting disproportionately coming from poorly performing hospitals.

There are reasons to be skeptical of the perceived wait time numbers from these surveys. One study (Davis and Heineke, 1998, supplemented by personal communication with Davis, 2008) that actually compared consumer-perceived waiting times for a service encounter in a fast-food restaurant with actual waiting times, as measured by a stopwatch, found that perceived waiting times were on average more than 25% higher than actual wait times. These results are likely to be quite conservative compared to those comparing actual and perceived wait times in the British studies, for a number of reasons. First, in the Davis and Heineke study, customers were asked about perceived wait times immediately after their customer service experience, while memory was fresh. By contrast, the two audit studies took one to three months after patients had visited A&E, creating significant memory problems. (Fieldwork for the 2003 study occurred in February 2003, and involved patients treated between November 2002 and January 2003; fieldwork for the 2004 study occurred between August and October 2004, and involved patients treated between June and August 2004; personal communications from Karen Hallt, Healthcare Commission, 2008). One might argue that poor memory might affect the standard deviation of responses, but not the mean. However, since underestimates are truncated at zero but there is no limit to overestimates, poor memory biases the mean perception upwards, as well as the percentage of patients reporting having waited over four hours. Second, compared to ordering at a fast-food restaurant, coming to an emergency room is far more emotion-laden, where one is anxious to be treated quickly, and a wait time of several hours will surely seem close to infinite to some. Finally, if one assumes that the patient perception numbers are correct and the official ones fraudulent, this implies a level of fraud so enormous that it is inconceivable it would have gone unnoticed or unexposed (by whistleblowers, if nobody else).

In 2004-5 the U.K. Audit Commission studied hospital data quality, including for A&E departments (Department of Health, 2005; Audit Commission, 2005). Auditors rated 137 trusts as having good audit trails for wait time data, and 19 as lacking them. Comparing computerized and manual records for a sample of cases at each trust, auditors found errors in 10.6% of cases. These results are of some use, but don't answer all questions about data quality. We don't know how large errors were, what proportion affected breaches of the four-hour target, and to what extent they are concentrated in ex ante poorly performing hospitals. (In the most extreme case that every error was an overestimate of wait time that resulted in a failure to report a true breach of the four-hour target, this would still not come near to accounting for the difference between the official and the patient perception data.) The report makes no suggestion of fraud as a source of any data-quality problems.

the period. Mean wait time also decreases[21] dramatically by one-third, from 2.50 hours to 1.88 hours; the percent of patients treated within two hours increases as well.[22]

Table 1 also presents descriptive statistics for our measures of effort substitution and gaming, all of which improve during our sample period. Death rates decrease 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. Wait times for orthopedic surgery decline dramatically, falling nearly in half. The percent of patients admitted to inpatient wards decreases slightly.

Hypothesis Testing

Table 2 reports results from a 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 (odd and even columns). For the moment, we focus on the results in odd-numbered columns. 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 dependent variable is the death rate, measured quarterly, and the independent variable is the change in wait-time performance (Δx_{ht}). The F-test for the joint significance of the time-period fixed effects is greater than 20 in all specifications, indicating significance far better than the 1% level; the similar test for joint significance of the trust fixed effects often yields values in the thousands, and so is extremely significant. The R-squared for the regressions all fall

[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 ANONYMIZED, 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.

between 0.65 and 0.86, with the marginal increase from the addition of the independent variable(s) of interest varying between 0.1 when the coefficient is highly significant and 0. Due to the many specifications we omit these statistics from Tables 2 and 3.

Hypothesis One states that improved wait times would be associated with lower quality of care. Instead, we find only 1 coefficient out of 12 that is significantly greater than 0, in the second panel of Column 1. For 7 specifications there is no significant effect. And in 4 specifications, there is a significant effect, but it goes in the direction opposite to that predicted by the hypothesis. As wait-time performance increases, the death rate decreases, although the magnitudes of the improvements are somewhat small. A rather large 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 average changes in performance over time; thus, this regression shows that hospitals which improve most along the measured dimension, in a given period, 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. The specifications using return visits as the alternative measure of quality also show no statistically significant effects, and the estimates are precise enough to rule out all but the smallest magnitudes of effort substitution effects.

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 column 1. 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 months, barely more than one one-hundredth of a standard deviation. The precision of our estimates thus allows us to rule out all but the smallest of dysfunctional effects in these data. Hypothesis Two is not confirmed.

Hypothesis Three states that treating more patients in fewer than 4 hours would be associated with an increase in mean wait time. Table 3 presents these results in columns 3 and 5. 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 actually decreases.[23] Similarly, most of the significant coefficients in regressions using the sub-two hour fraction are negative. Thus, Hypothesis Three is not confirmed; instead there is again evidence, here quite strong, that is the opposite is occurring: fewer waits longer than four hours are associated with lower mean wait times and a higher fraction of patients treated in under two hours.

[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.

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]

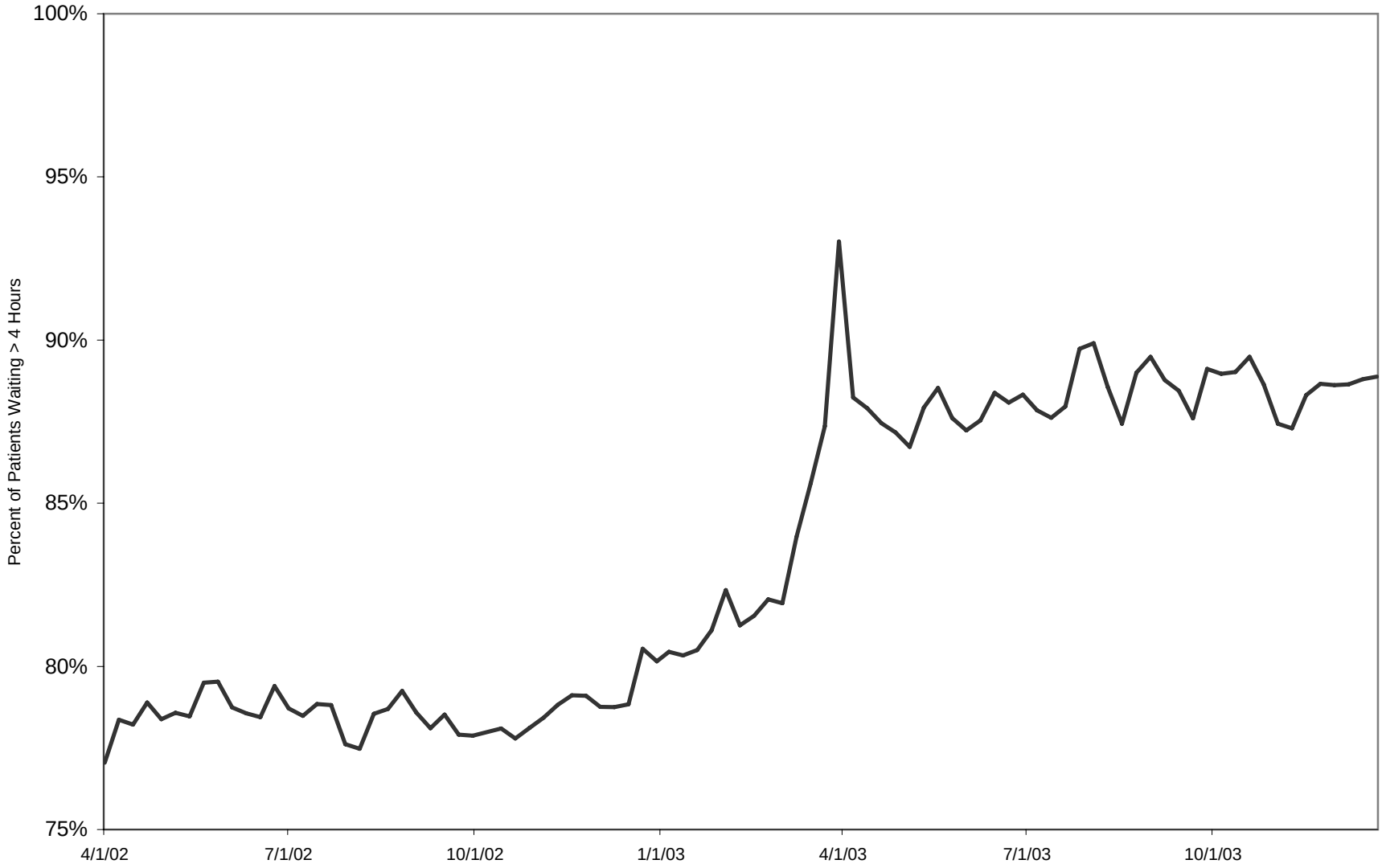
performance improvement occurring during that week persisted thereafter.

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]

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

Figure 2: Emergency Room Waiting Times



Hypothesis Five states that improved wait times would be associated with an increase in admissions into inpatient wards. Results from regressions testing this hypothesis are in column 7 of Table 3. Once again, 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.[26]

[26] Mayhew and Smith (2008) present an analysis of A&E waiting times in which they fit a queuing model onto observed data from 2002. This analysis concludes that meeting the "target would not have been possible without some form of patient re-designation or re-labeling taking place" (by which they mean moving the patient into another part of the hospital such as an inpatient ward), since their model generates the result that average waiting times in 2006 would have had to have fallen by an extreme, unrealistic (and counterfactual) amount to meet the 98% target for 4-hour wait without such re-designation. The conclusion Mayhew and Smith draw is driven by their choice of an exponentially distributed treatment model. The exponential distribution has both a large tail and only one parameter that can vary. Since there is only one parameter, one cannot move the "hump" and the "tail" separately; they must move together. In particular, the exponential distribution simply scales out proportionally – that is, reducing the 98th percentile of wait times from 6 hours down to 4 hours (for instance, as many trusts did) necessarily entails reducing every other percentile to two-thirds its previous value. Thus, reducing the "tail" of the distribution requires, a proportionate reduction in the "hump." In our setting, the assumption of an exponential distribution necessitates that one cannot reduce extremely long waits without also reducing all waits. Thus, reducing the longest treatment times entails proportional reductions in all waiting times by the assumptions of the model, which in turn generates Mayhew and Smith's claim that there would need to be extreme reductions in average waiting times to meet the target that could only be met by "re-designation."This is not an empirical result; it is a one that is logically entailed by the model they assume. Mayhew and Smith provide no empirical evidence for the claim that the target could be met only by "re-designation," and our analysis using actual inpatient admission data (which Mayhew and Smith do not use) does not support their theory. (It may be noted that they assume the form of "re-designation" would be assignment to newly established "medical assessment units"; such assignment is recorded in the data as admission to inpatient wards.)
The problem with the Mayhew and Smith analysis is that there are many models likely to fit the 2002 data equally well, and some of these models would yield very different results from theirs. Indeed, they note (p. 13) that "we make no claim that [an exponential model] is the best possible model of its type there is." For instance, any model that allows hospitals to reduce the "tail" of wait times without moving the "hump" – that is, concentrate on treating long wait-time cases faster, certainly an empirically plausible strategy for a hospital – would allow hospitals to meet the 4-hour target without shortening other waiting times by nearly as much Mayhew and Smith suggest. As a simple example, suppose that the time for treatment comprises two parts: time for actual treatment and time wasted by inefficiency. Suppose that the actual treatment time is relatively compact – for simplicity, suppose treatment is uniformly distributed between 1 and 4 hours, reflecting the severity of the ailment. Suppose that wasted time is usually low but can, in rare cases, be very high – for simplicity again, suppose that the time wasted is 0 hours with 90% probability and 5 hours with 10% chance. In this simple model, the expected wait time will be 3 hours (2.5 for treatment and 0.5 for inefficiency), and the unluckiest 2% of patients will wait longer than 8.4 hours. Now suppose that hospitals eliminate inefficiency, with no change in the time for real treatment. Now the average treatment time is 2.5 hours, and all patients are treated in 4 hours. Most patients have not seen decreased wait times (only those few previously hit by the 5 hour time inefficiency). In this model, trusts have reduced the 98th percentile from 8.4 to 3.94 hours while only reducing the average treatment time by 0.5 hours. In this model, the average need fall much less than proportionally relative to the tail.
Given the widely varying conclusions that would have been generated from other models that would have fit the data equally well, the approach Mayhew and Smith take of using one assumed model to draw conclusions is not a fruitful one.

ALTERNATIVE EXPLANATIONS AND ROBUSTNESS CHECKS

The primary source of potential bias in the results presented in Tables 2 and 3 is omitted variables bias (OVB). For instance, suppose that the average severity of patients visiting the A&E varies from week to week. As the seriousness of illness decreased, the percent of patients treated within four hours would increase, as would a number of our alternative quality measures, producing a negative bias that could explain some of our results.

We estimate two alternative specifications in order to investigate the potential for OVB. We first 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 to increases in the measured statistic between incentivized and non-incentivized periods. Any potential omitted variables bias would be present in all periods, but the effort substitution or gaming effects should appear only when incentives are in place. Thus, if $\gamma>0$, there is evidence of effort substitution or gaming that is free from bias. As before, we cluster standard errors by trust. These results appear in even-numbered columns of Tables 2 and 3.

Of 28 estimated coefficients (using all 4 versions of the dependent variables and 7 measures of alternative performance), there are three coefficients that are significantly

greater than 0. But all are of small economic magnitude. For instance, there are two such significant coefficients in the second panel of columns 4 and 6 in Table 3. Relative to a similar move in periods without incentives, increasing performance by one percentage point increases the fraction of patients waiting more that two hours by 0.275 percentage points. This effect is about 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. Furthermore, the coefficients that are not statistically different from 0 are estimated precisely enough to rule out all but small magnitudes of effort substitution or gaming. Our estimates of equation (2) thus also fail to confirm any of our hypotheses about effort substitution or gaming.

We also present instrumental variable (IV) estimates in Table 4, in which we re-estimate the main specifications in Tables 2 and 3 using the timing of the incentive periods as an instrument for performance or performance improvement. This instrument is excellent for performance improvement, but since performance both increases during incentive periods *and* remains high after incentives have ended, incentive periods are a poor instrument for the level of performance. Performance is lower before the incentive periods but higher after, so, on net, there is no difference. Instead, to estimate the IV specification using *Performance* and *Log(Performance)* as the right-hand-side variables, we restrict the sample to the quarter before and the quarter including the implementation of incentives.[27] Since the instrument does not vary across trusts within a time period, we can no longer use quarter fixed effects, which would subsume the instrument. Instead, we use a polynomial in time (linear for the level regressions, quadratic for the difference

[27] Specifically, we use 2002Q4, 2003Q1, 2003Q4, 2004Q1 in calendar time.

regressions). Additionally, we include quarter-of-year fixed effects in the difference regressions to control for the seasonality present in a number of our variables.[28]

Results from these regressions appear in Table 4. The first-stage regression coefficients (regressing the relevant performance measure on a dummy for "incentives active" as the instrument) appear at the left of the table. The instrument is highly significant in all cases, with t-statistics of 5.03, 4.44, 3.58, and 5.62, respectively. The remaining 12 columns provide the IV estimates in even-numbered columns and the analogue OLS regression in odd-numbered columns. (The OLS results in this table are different from those in Tables 2 and 3 because the time-controls are different – we

TABLE 4 ABOUT HERE

provide them as a basis for comparison). Of 24 different IV coefficients, only 4 are greater than zero, and none is significantly so. On the other hand, 12 are significantly less than zero. The IV estimates provide yet further evidence that is inconsistent with effort substitution and gaming and suggests that OVB is not the driver of our results.

A second source of bias may be reverse causality from the alternative measures of performance back onto the measured statistic. For instance, suppose that different forms of effort are substitutes, as is typically the case. If a trust were exogenously to reduce the effort directed towards preventing follow-up visits, it would be natural to increase the effort directed to reducing wait times below four hours. The IV results in Table 4 weigh against this possibility, though, unless the incentive period itself directly increased performance on the alternative measures, which then created an impact on the targeted statistic. Such an indirect effort seems unlikely. Furthermore, our results are, at some

[28] In the level specifications, the time trend essentially functions as a fixed effect for "early quarters" and "late quarters." Since each sub-period includes one Q4 and one Q1, we need not worry about seasonality of variables.

Table 4: Testing for Effort Substitution: Instrumental Variables Estimates

Explanatory Variables:	*First Stage Coefficient*	*Dependent Variable:* *Surgery Wait Times* (1)	(2)	*"Mean" Wait Time* (3)	(4)	*% Waits > 2 Hours* (5)	(6)	*Hospital Admits* (7)	(8)	*Deaths* (9)	(10)	*Follow-Ups* (11)	(12)
Performance	0.012**	-0.540	-20.99**	-3.138**	-4.092**	-0.646**	-1.249**	0.079*	0.050	-0.700**	-2.732**	0.037	-0.056
	(0.002)	(0.468)	(4.751)	(0.285)	(1.156)	(0.101)	(0.409)	(0.037)	(0.162)	(0.230)	(1.022)	(0.027)	(0.108)
ΔPerformance	0.009**	0.092	-17.53**	-1.106**	-0.230	-0.301**	-0.279	-0.003	0.381	0.125	-2.332*	-0.009	-0.339
	(0.002)	(0.276)	(4.814)	(0.170)	(1.191)	(0.050)	(0.372)	(0.025)	(0.257)	(0.168)	(0.970)	(0.014)	(0.224)
Log Performance	0.062**	-0.079	-4.106**	-0.392**	-0.747**	-0.084**	-0.228**	-0.011	0.010	0.094**	-0.535*	0.004	-0.011
	(0.017)	(0.075)	(1.240)	(0.051)	(0.245)	(0.015)	(0.084)	(0.006)	(0.032)	(0.029)	(0.218)	(0.004)	(0.022)
ΔLog Performance	0.112**	0.021	-1.396**	-0.096**	-0.028	-0.028**	-0.034	0.001	-0.034	0.022*	-0.186*	-0.001	-0.030
	(0.020)	(0.018)	(0.336)	(0.008)	(0.143)	(0.002)	(0.045)	(0.001)	(0.022)	(0.010)	(0.079)	(0.001)	(0.020)
Regression Type	FS	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Trust Effects?	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Time Controls?	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
N	608 / 2291	608 / 2290		595 / 1973		595 / 1973		608 / 1984		608 / 2291		603 / 1978	

Columns alternate OLS and IV specifications, where we instrument for the relevant measure of performance with the implementation of incentives. In each of the four first stage regressions, the instrument is highly significant with a t-stat of 5.03, 4.44, 3.58, and 5.62, respectively. The number of observations records two sample sizes: that for the level regressions, and that for the difference regressions, respectively. The level regressions use data only from the period before and after incentive implementation, that is 2002Q4-2003Q1 and 2003Q4-2004Q1. The difference regression include all quarters between 2002Q4 and 2006Q4, where data is available. Level regressions include a time trend, difference regressions include a quadratic funciton in time. Statistical significance is denoted with the system: * 5%, ** 1%. Standard errors are clustered at the trust level.

level, invariant to the direction of causality; if an increase in output 1 causes an increase in output 2 because they are complements, an increase in output 2 would of course also increase output 1. Thus, regardless of the direction of causation, our finding that targeted and alternative measures of performance covary positively is inconsistent with a model in which the two are substitutes, as must be the case in a model of effort substitution.

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.

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 do not claim that dysfunctional effects never occur; we note only that we find no evidence in this case -- although there were many claims, based on theory or anecdotal evidence, that they had -- and we seek to understand why they did not. 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, 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 (ANONYMIZED, 2008) 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 developing better procedures to aid

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.[29]

A second managerial behavior is to adapt measures 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 (Cullen and Reback, 2006).[30] An earlier iteration of efforts in the U.K. to reduce A&E wait times used commencement rather than completion of treatment as the target, producing an introduction of "hello nurses" who greeted the patient and allowed the claim treatment had begun; the treatment completion target discussed in this paper represented an improvement over the earlier one. Skeptics will see this as an example of cascading regulation (Zeckhauser, 1979), ultimately futile; 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 evolution of rules over time in response to learning new information (March, Schulz, and Zhou, 2000).

[29] 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.

[30] 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."

A third managerial behavior can be to harness the 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 organization's mission goals.[31] 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 performance improvement and the percentage of students excluded from the test. The difference might be due to different levels, or different levels of managerial mobilization, 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, but it may reduce its magnitude.

Limitations

Since wait-time performance is a predictor variable in these models, we need to ask whether, or under what circumstances, data falsification would affect these results. The intuition is that if poorly performing hospitals appear to be performing better than they are because of data falsification, this will bias coefficients towards zero, in the direction of our results, and that this might be why we fail to find significant distortionary effects. We are skeptical of suggestions of major data falsification, but we note that such falsification, even were it hypothetically to exist, would not necessarily affect our results.

[31] 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.

Even were wait-time data falsification of a large magnitude and concentrated among poorly performing hospitals, if it is constant over the time period of these data, its impact would be picked up through the hospital fixed effects variable. Only if falsification were of a large magnitude and increasing over time in just those hospitals that improved the most would falsification bias coefficients towards zero. The problem of biasing coefficients towards zero would also occur in the case of random measurement error in any of the variables in these models, or for significant data falsification that was randomly distributed across hospitals and also constant across time periods, which both introduce noise and thus bias coefficients toward zero. It should be noted that, if any of this were occurring, then for those hypotheses where we find significant results in the direction opposite from that predicted by the distortion stories, our results would in reality be even stronger than those we have reported.[32]

We also wish to make clear what we have sought and not sought to do in this paper. We have focused on whether the target regime produced distortionary responses. We have also noted that both attainment of the four-hour target and mean wait time very dramatically improved between 2003 and 2006. We cannot claim based on the data in this paper that all this improvement is due to attention to the A&E target or that the social benefits of the improvement were greater than its financial costs. As noted earlier, hospital budgets increased during this period, so more resources were available for A&E departments, though we have noted that in the public sector budget increases frequently fail to generate performance improvements. We should also note that, if one believes the

[32] It is certainly possible that falsification has become less of a problem over time, not more, due to greater organizational and auditor attention to the issue. If falsification were concentrated to the poorest-performing hospitals and got better over time, then the results we report would become conservative. If data quality were improving over time in all hospitals, this would be captured by our time period fixed effects variable.

patient perception surveys discussed (and critiqued) earlier, or widespread data falsification, the improvements noted here would be exaggerated, and the social benefits of the intervention would be less.

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.[33]

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.[34] The same goes for gaming. Blau

[33] 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?"

[34] 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

(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.

---

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 wait 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. 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.

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.

Hanushek, Eric A. 1996. School resources and student performance. Pp. 43 – 72 in *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, edited by Gary Burtles. Washington, DC: Brookings Institution Press.

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.

Healthcare Commission. 2005 Accident and Emergency. Available at http://www.healthcarecommission.org.uk/_db/_documents/04019296.pdf

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

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. Wait 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.