Does ‘naming and shaming’ work for schools and hospitals? Lessons from natural experiments following devolution in England and Wales

Gwyn Bevan and Deborah Wilson

Prior to devolution in 1999, governance of schools and hospitals in England and Wales was similar. After devolution, the funding and organization continued to be similar, but the two governments adopted different policies in the pursuit of common objectives. This paper reports the results of two ‘natural experiments’ which compare outcomes in the two countries before and after these policy changes. The governance model of ‘trust and altruism’ resulted in worse reported performance in Wales as compared with England on what were each government’s key objectives. We argue that ‘naming and shaming’ worked in England, as compared with Wales, resulting in improved examination performance and eliminating the endemic problem of long waiting times.

Keywords: Governance; hospitals; league tables; performance measurement; ranking.

In 1999, the Labour government devolved powers to the governments of the different countries in the UK. The Welsh Assembly Government sought to create ‘clear red water’ in its policies for Wales with differences from England in governing schools and hospitals (Greer, 2009). Prior to devolution, for both countries, there was the same legislation (unlike Scotland and Northern Ireland), and similar institutional arrangements, governance, and levels of funding. Schools were largely run by local authorities; health services were largely run through organizations accountable to central governments. Following devolution, similarities in levels of funding and institutional arrangements continued. We report in this paper studies of two ‘natural experiments’ in which the governments of both countries pursued the same policy objectives but with different models of governance: for schools, where the government in Wales stopped publication of school league tables at age 16, which continued in England; and, for hospitals, where the government in England implemented a similar policy of ‘naming and shaming’ by reporting hospital performance in ‘star ratings’, which was not followed in Wales.

The paper outlines four models of governance, describes our rigorous econometric studies of these two natural experiments and discusses what we can learn from their results.

Models of governance

This section outlines four models of governance that are similar to those outlined by key advisers to the Blair government’s reforms of public services: Le Grand (2007) and Barber (2007). We outline each model’s characteristics and conditions for effectiveness with reference to three sets of concepts: ‘knights’ and ‘knaves’ (Le Grand, 2003), ‘econs’ and ‘humans’ (Thaler and Sunstein, 2008), and prospect theory (Tversky and Kahneman, 1991). The different models rely on some form of summary measurement of (comparative) performance, of ‘quality’. Models that generate high-powered incentives based on the limited measures of quality that are available run into problems of gaming (Wilson, 2010). In practice different models may use similar information, and policies often use a mix of these models, as we show in our two case studies.

Trust and altruism (T&A)

The T&A model assumes that those who deliver public services are purely driven by ‘knightly’ motives of altruism and hence all that is needed for better performance is comparative information (Le Grand, 2003, 2007). This model ignores ‘knavish’ motives of self-interest and hence assumes that there is no need for external incentives to overcome organizational inertia, both to understand how performance needs to change to deliver better outcomes, and to
implement the necessary changes (Berwick et al., 2003). Furthermore, as failure in this model can only be due to difficult external circumstances, it ought to result in extra resources. The consequences are perverse incentives that run counter to prospect theory (see below). This model lacks theoretical justification, has low monitoring costs, is popular with professionals (Le Grand, 2007), is common for public services, and indeed seems to be the default model for the countries of the UK.

Hierarchy and targets (H&T)

We define the H&T model as having a limited set of public targets that clearly signal priorities, with specified rewards for ‘successful’ organizations and sanctions directed at those responsible for running ‘failing’ organizations. Thaler and Sunstein (2008) argue that assumptions in elementary microeconomic models of individual behaviour (as ‘econs’) differs from that of real individuals (as ‘humans’). The H&T model assumes providers are ‘econs’ and respond to clear economic incentives. Prospect theory (Tversky and Kahneman, 1991) tells us that ‘humans’ feel losses much more keenly than gains of equivalent magnitude (‘econs’ are assumed to experience symmetrical losses and gains in their utilities); and hence that, of these incentives, it will be the sanctions that will have the stronger impact. This model is also known as ‘command and control’ (Barber, 2007; Le Grand, 2007). Barber, based on his own experience, argues that this model ‘done well can rapidly shift a service from, ‘awful to ‘adequate’’ (Barber, 2007, p. 335). The H&T model imposes external incentives by strong performance management, has monitoring costs and is unpopular with professionals (Bevan and Hood, 2006; Le Grand, 2007).

Choice and competition (C&C)

Le Grand (2007) argues that the C&C model, in the form of quasi-markets for public services, can be designed to generate incentives that appeal to both ‘knightly’ and ‘knasivsh’ motives in providers, and enable users of public services to move from being grateful recipients (‘pawns’) to empowered consumers exercising choice (‘queens’). The C&C model assumes users are ‘econs’ and choose better performing providers based on available information on quality. This model creates external incentives via a quasi-market system in which users have a choice of providers and money follows their choices (the pupil or patient). Barber (2007, p. 335) advocates this model as offering the means to high-performing public services. He quotes Blair’s antipathy to ‘flogging’ by the H&T model and his belief that ‘Innovation should come from self-sustaining systems’. As Barber (2007) and Le Grand (2007) recognize, effective quasi-markets depend on a series of conditions. These include supply-side flexibility, which is not required by the other models we outline. Some ‘failing’ schools have been closed and new schools created (sometimes on the same premises: see Ball, 2008); but even closing hospital departments typically evokes strong opposition and there are formidable barriers to entry for new hospitals and hospital services (Breereton and Vasoodaven, 2010; Tuohy, 1999). Quasi-markets have high transaction costs, but are popular with governments because pressure on poor performance comes from an ‘invisible hand’ (Le Grand, 2007) and they promise to have more potential to respond to users’ needs than centrally-driven systems of hierarchy and targets and reputation.

Transparent public ranking (TPR)

Barber (2007, p. 336) advocates a model of ‘devolution and transparency’ in contexts where neither the H&T nor C&C models would be effective. He gives as an example the New York City Police Department with devolution of operational responsibility and resources to, and publication of crime results by, each precinct. Our TPR model is of this type and is based on four criteria that Hibbard et al. (2003) identify for a system of performance measurement to be effective. It has to be: a rankingsystem; published and widely disseminated; easily understood by the public (so that they can see which providers are performing well and poorly); and followed up by future reports (that show whether performance has improved or not). The TPR model assumes that providers are ‘humans’ and respond to threats to their reputation, which may, of course impact on their careers and hence also appeal to ‘econs’. Prospect theory tells us that the sanctions of ‘naming and shaming’ providers that perform poorly in the rankings at any time, or perform worse over time, will have a stronger impact than ‘naming and framing’ that celebrates the success of those who perform well. This model has monitoring costs and is unpopular with professionals.

The ‘natural experiment’ of stopping publication of school league tables in Wales from 2001

The changing mix of models of school governance in England and Wales

Notoriously, early editions of the Encyclopaedia Britannica had in the index the entry ‘For Wales see England’ (Robbins, 2001), which was seen as
a good description of the policies for schools in Wales prior to devolution. As schools were seen to be underperforming under the T&A model in the 1980s, they were chosen by the UK Conservative government as one of the first areas for quasi-market reforms, which involved a mix of the C&C, TPR and H&T models. The Education Reform Act 1988 introduced local management of schools, over-lapping catchment areas and open enrolment of pupils, with money following the pupil. From 1992 there was annual publication of school performance tables for secondary schools, ranked by local and national media, based on examination results at age 16 to inform parental choice of school in the C&C model (Wilson and Pielbaga, 2008, discuss how the league tables have since evolved). Reports of school inspections were published, by the Office for Standards in Education (Ofsted) in England, and Her Majesty’s Inspectorate for Education and Training in Wales (Estyn) in Wales. Regular public reporting of school performance information also brings pressure to improve performance through the TPR model. The H&T model was applied to ‘failing’ schools (as measured by league table position, sometimes relative to centrally set floor targets) and/or inspection results, but was less well-defined than for hospitals (see below). For English schools the mix of interventions included: additional resources, increased monitoring, school takeover, re-opening and/or closure (Ball, 2008; Allen and Burgess, 2012). After the 1997 election, the Labour government maintained this mix of policies in England and Wales.

Following devolution, the Welsh Assembly took responsibility for education spending and policy. The National Union of Teachers in Wales called for abolishing the publication of school league tables (BBC, 1999). After a consultation exercise, the Welsh Assembly Government announced in July 2001 that it would stop the publication of school league tables with immediate effect. There were no other major policy differences between England and Wales: the national curriculum, the school inspection regime, the setting and marking of the high-stakes examinations at the end of compulsory schooling at age 16 all remained the same. So, in Wales, the TPR model was abandoned and the C&C model continued, but without comparative information from school league tables and with recognition of its limited impact in its many rural localities (Reynolds, 2008).

Study design
Isolating the impact of any one element of a system of governance is difficult for two reasons (Figlio and Ladd, 2008): lack of an adequate control group for the counterfactual, and the typical introduction of a multifaceted reform at the same time. The study we outline here, by Burgess et al. (2010), avoids these two major problems by exploiting a ‘natural experiment’ to evaluate the effect of abolition of school league tables in Wales: students in English schools provide a ready-made control group, and the rest of the education system remained essentially unchanged across the two countries. Using administrative datasets from school censuses of all state schools in England and Wales, the study implemented a classic difference-in-difference (D-in-D) analysis comparing outcomes in England and Wales, before and after the 2001 reform. Outcomes were measured in two ways: the proportion of pupils achieving five or more ‘good’ GCSE grades (the headline figure from the league tables, see Wilson et al., 2006), and a measure of value added: this is the mean GCSE point score, controlling for performance in Key Stage 3 tests at age 14 (see Burgess et al., 2010). This enabled analysis of the extent to which any changes in outcomes were true improvements in school effectiveness rather than a gaming response to the headline figure.

Study results
Figure 1 plots the country-year averages of mean GCSE point score in English and Welsh schools. The vertical line indicates the timing of the policy change. (The policy announcement was in July 2001, just after GCSEs had been taken. Hence 2002 GCSEs were the first to be taken under the knowledge that the results would not be published.) The figure illustrates

Figure 1. School mean GCSE points score in England and Wales over time.
Note: Connected points indicate the before and after periods in our analysis.
Source: Burgess et al. (2010).
that before the change performance in Wales was improving more than in England, but afterwards this was reversed. A simple D-in-D confirms this result in table 1. Results in England did improve faster than in Wales after the reform, with the difference estimated to be 2.2 GCSE points. (This is equivalent to 0.27 of a school-level standard deviation; a substantial effect.)

These are raw data, however, so the observed differences in outcomes may be the result of a range of factors alongside the reform itself. In order to isolate the extent to which the difference in the differences are due to the abolition of school league tables, a full econometric D-in-D analysis was undertaken which controlled for school characteristics, students’ prior attainment, schools’ expenditure, level of competition. The analysis was undertaken using two samples of schools: the results reported here come from the matched sample, whereby Welsh schools were matched with similar English schools on the basis of a range of observable characteristics.

Table 2 confirms that there was a negative impact on school performance in Wales relative to England following the abolition of league tables in 2001, by almost two GCSE grades per student per year. The average proportion of students gaining at least five good GCSE grades in Wales also fell, relative to England, by approximately 3.5%. Further analyses showed that the top quartile of Welsh schools appeared unaffected by the reform. It was the lower 75% of schools—as measured by student prior attainment and by poverty—that were negatively affected, with the poorest and lowest average ability schools falling behind the most. These results were not driven by English schools gaming the league tables after 2001: the estimated effects were similar in size for both published and unpublished performance measures, and were confirmed by the differences in results for English and Welsh schools in the independent PISA tests. The conclusion of the study was that it was not C&C that was driving the impact of the policy change. The results did not vary by the degree of competition schools faced, and the largely rural nature of the matched sample of schools illustrated the low level of potential choice in Wales. We discuss this further below.

The ‘natural experiment’ of the consequence of ‘star rating’ National Health Service (NHS) hospitals from 2000 to 2005

The changing mix of models of hospital governance in England and Wales

The Thatcher government moved away from the T&A model to one of C&C for NHS hospitals in a quasi-market across the UK. This was similar to that for schools: ‘money followed the patient’ and NHS hospitals were separated from their health authorities to become independent NHS trusts. After winning the 1997 general election, the Labour government abandoned the C&C model for T&A (Bevan and Robinson, 2005). Prior to devolution, the NHS in Wales had similar policies and practice to those in England, being perceived as ‘forming an adjunct to the English health service’ (McClelland, 2002, p. 325). Following devolution, the government in Wales continued with the model of T&A (Auditor General for Wales, 2005; Bevan, 2010). In 2000, in response to a perceived crisis of ‘underfunding’ of the NHS, the Labour government decided on a policy of sustained real increases in funding, of about 5% a year for five years for England (Smee, 2005), which also applied to the NHS in Wales (Connolly et al., 2011). The government in England abandoned the model of T&A as it

### Table 1. Simple difference-in-differences.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean GCSE score</td>
<td>38.68</td>
<td>44.46</td>
<td>5.78</td>
<td>38.64</td>
<td>42.22</td>
<td>3.58</td>
<td>-2.2</td>
</tr>
</tbody>
</table>


### Table 2. Full difference-in-difference results.

<table>
<thead>
<tr>
<th>Treatment effect</th>
<th>School mean GCSE score</th>
<th>School percent 5 A*-C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.923*** (0.546)</td>
<td>3.432*** (0.950)</td>
</tr>
</tbody>
</table>

sought improvements in NHS performance through the policies of TPR and H&T, which applied from 2000 until 2005, through the regime of annual ‘star ratings’ (Bevan, 2010).

‘Star ratings’ gave NHS trusts a score from zero to three stars based on performance against three sets of data. For 2002/03, for example, these were: a small number of ‘key targets’; a larger set of targets (dominated by targets for waiting times) and indicators in a ‘balanced scorecard’ (including more targets for waiting times, hospital mortality rates and results from surveys of patient satisfaction); and results from inspections of their systems of clinical governance to assure and improve quality of care (Commission for Health Improvement, 2003). This regime sought to change the culture of the NHS from one in which a hospital’s failure to achieve targets for waiting times was rewarded with extra funding, to one in which such failure resulted in sanctions, and success in rewards. In line with prospect theory, the sanctions for failure attracted more attention than the rewards of ‘earned autonomy’ for success. The ‘targets and terror’ element of H&T (officially described as one of ‘franchising’) featured prominently in the first two years in which it was made clear that chief executives of zero-rated hospitals were at risk of being sacked if they had been in post long enough and failed to produce a convincing strategy to turn round their hospitals’ performance (Bevan and Hood, 2006). In the first year, the 12 hospitals that were zero-rated were publicly ‘named and shamed’ by being called the ‘dirty dozen’ and six of their chief executives were sacked.

As the Welsh Assembly Government was looking to stop publication of school league tables, they decided not to introduce an analogous regime to that of star ratings for the NHS in Wales (see below). In England, in 2002, the government signalled a shift in emphasis from H&T and TPR to C&C, which mainly took effect from 2006, after the end of the ‘star ratings’ regime in 2005 (Bevan and Robinson, 2005; Dixon and Mays, 2011). In Wales, however, the government moved in the opposite direction and removed the purchaser/provider split with health boards taking over the running of NHS trusts (Welsh Assembly Government, 2009).

Study design
The study we outline here, by Besley et al. (2009), used the method of D-in-D to compare England and Wales on waiting time performance for elective day case or ordinary admission. There was no change in the definitions of cases on waiting lists over this period, but in Wales all referrals were included (whatever the source), whereas in England only referrals from medical and dental general practitioners were included (Auditor General for Wales, 2005 pp. 50–53).

The data analysed were the number of patients in seven three-monthly waiting bands (zero to three months, three to six months etc., with the seventh band being patients waiting more than 18 months). These data were obtained for each NHS trust in Wales and England on the last day of each financial quarter for seven years: from 1999/2000 (before the ‘star rating’ regime was introduced) to 2005/2006 (a year after the end of that regime). The effect of each waiting time target in England, with Wales as the control, was estimated by running for each waiting band a D-in-D model with a variable that took the value of one for England for the year and time that band was the target and zero otherwise.

The targets that applied in England, for each year (beginning in April and ending in March) for patients on waiting lists were as follows:

- 18 month target for 2000/01: no patients wait longer than 18 months at the year end.
- 15 month target for 2001/02: no patients wait longer than 18 months at each quarter, and fewer patients wait for more than 15 months at the year end (than at end of March 2001).
- 12 month target for 2002/03: no patients wait longer than 15 months at each quarter and no patients wait longer than 12 months at the year end.
- Nine month target for 2003/04: no patients wait longer than 15 months at each quarter and no patients wait longer than nine months at the year end.
- Nine month target for 2004/05: no patients wait longer than nine months at each quarter.

The Welsh target regime was within the model of T&A: there were so many targets that it was unclear which mattered, and failure to achieve waiting time targets was not publicised and was perceived to result in extra funding (Auditor General for Wales, 2005).

Study results
In June 1999, prior to the implementation of the target regime, the mean total number of patients waiting more than six months for an average NHS trust in England and Wales was similar (about 1800 in Wales and 1700 in England). Those waiting on lists in Wales tended
to wait longer, however. In particular over 200 patients in Wales had been waiting more than 18 months compared to none in England. The raw data show that, when each successive target came into force in the English NHS, there were negligible numbers of patients waiting longer than each target. Figure 2 shows the divergence in performance between England and Wales from 2001 for the numbers per thousand waiting for more than six months in each country.

A difficulty in using the raw data to assess the impacts of the ‘star rating’ regime in England in comparison with Wales is that the performance in England was better than in Wales before the introduction of that regime. The benefit of the D-in-D model is that it estimates the impacts of that regime on the differences in performance that existed prior to its introduction. The estimates from the D-in-D model, which were statistically significant, show the effects of the targets to be as follows:

• 18 month target for 2000/01: reduced the numbers waiting more than 18 and 12 to 15 months, but increased those waiting less than six to nine months, and three to six months.
• 15 month target for 2001/02: reduced the numbers waiting more than 15 to 18 and 12 to 15 months, but increased those waiting six to nine, and three to six months.
• 12 month target for 2002/03: reduced the numbers waiting more than 12 to 15 months and nine to 12 months, but increased those waiting three to six months.
• Nine month target for 2003/04: reduced the numbers waiting more nine to 12 months and six to nine months, but increased those waiting less than three months.

Hence NHS trusts in England responded by ‘tail gunning’: i.e. each year, focusing on eliminating the long waits that put them at risk of missing the targets for that and the following year, at the expense of those waiting much shorter times than either target.

Discussion

We summarize here the results of the D-in-D analyses of two ‘natural experiments’, which we have reported, of differences in policies for schools and hospitals in England and Wales. We consider why these differences emerged following devolution, the relevance of our findings from a decade of generous funding of schools and hospitals for an austere future, and make a plea for better comparative data.

The two studies we report, in effect, compare the shift for schools in Wales from the TPR model to T&A from 2001, while the TPR model continued in England; and the shift for hospitals in England from 2000, from T&A to TPR and H&T, while the T&A model continued in Wales. The consistent finding is that the T&A model resulted in worse reported performance in Wales as compared with England on what were each government’s key objectives of improving examination performance at age 16 and reducing long hospital waiting times. This is strong evidence given the closeness of the systems prior to these changes in models of governance, and the similarities in funding and organization before and after those changes. We are aware of evidence of ‘gaming’ in response to the publication of school league tables and ‘star ratings’ and thus that these systems did turn some ‘knights’ into ‘knaves’. Kelman and Friedman (2009) point out that scholarly writing has tended to focus on such dysfunctional consequences, and rightly argue that the fundamental question in evaluating the use of measures of performance is whether their consequential benefits, given the dysfunctional effects from gaming, were better or worse than the counterfactual. Our assessment is that counterfactual evidence from Wales suggests that the benefits of these systems did indeed outweigh their dysfunctional consequences. Furthermore, ceterus paribus, we would expect schools and hospitals in Wales to be better governed on the grounds of its relatively small scale (with a population of three million compared with that of 53 million in England). So we argue that our findings show that the T&A model not only lacks the theoretical
justification we identified earlier, but has been found wanting in our (and other) empirical studies.

Although, after devolution, the model of governance in Wales offers the counterfactual of T&A for schools and hospitals to the TPR model, in England a mix of models were used. For schools, the C&C model in principle applied throughout in England and Wales, and indeed was the original justification for the publication of school league tables. For hospitals in England, the ‘star rating’ regime combined the models of TPR and H&T (‘targets and terror’). Our assessment is, however, that the key driver for improved performance in each case came from the reputation effects of ‘naming and shaming’ in the TPR model. For schools, as we explained above, the matched sample was located largely in rural areas, where there was limited choice and hence impact from the C&C model. In ‘star ratings’, the Department of Health regime of ‘targets and terror’ was not continued after the second year; and it seemed that once this had strongly signalled the change in the culture away from one of rewarding failure, the Department of Health relied on the TPR model. So our conclusion from these two natural experiments is that ‘naming and shaming’ did work in England, as compared with Wales, resulting in improved examination performance and eliminating the endemic problem of long hospital waiting times.

As Hibbard et al. (2003) point out, the TPR model is effective in generating incentives for improved performance by inflicting reputational damage through ranking, but such pressure is unpopular with those who are ‘named and shamed’. Hence governments vulnerable to producer capture will seek alternatives. Barber (2007) argued that the Prime Minister’s Delivery Unit had the advantage of being located in the Prime Minister’s Office and hence was much less vulnerable to producer capture than, for example, the departments responsible for providers. Hood has argued the greater scale in England creates a greater degree of relational distance (Hood et al., 1999) between governments and providers of public services than in Scotland and Wales, where there are ‘smaller societies with more tightly-linked and overlapping political and social elites’ (Hood 2007, p. 97). McClelland (2002, p. 328) identified within the NHS in Wales, for example, ‘the existence of a closely integrated policy community’. This may explain why the government in Wales after devolution abandoned publication of school league tables and did not introduce a regime similar to that of ‘star ratings’ for hospitals. Hood and Dixon (2010), however, found little clear evidence, in terms of direct electoral benefits, for the tougher target regime in England, as compared with that in Wales (or Scotland).

Looking to the future, we recognize that both our ‘natural experiments’ took place in a period of significant real terms budget increases in the UK for both education and health. As TPR is less costly than C&C it may be favoured for securing performance improvement in the current era of fiscal austerity. But the unpopularity of TPR with providers may mean that, for such systems to be politically sustainable as a long-term policy, they may be better applied (and disguised?) alongside a C&C rather than a H&T model. This in part explains why, from 2006, the Labour government emphasised the C&C model (Bevan and Robinson, 2005) and the Coalition government abandoned TPR for controversial policies of C&C (Secretary of State for Health, 2010; Timmins, 2012). A form of ‘star ratings’ for health and social care might, however, be reintroduced for England: as this paper was going to press, the Nuffield Trust has produced a report for the Secretary of State for Health in England on the pros and cons of rating systems for providers of health and social care (Nuffield Trust, 2013).

This paper has shown the potential for quantitative, cross-national comparative work to inform policy debate and practice, but researchers can only realize that potential with good comparable data. We thus conclude our contribution to this PMM theme with a plea to those with responsibilities for public services in the UK to require its governments to produce such data (Connolly et al., 2011; Wilson, 2011).

References
Berwick, D. M., James, B. and Coye, M. J. (2003), Connections between quality measurement and improvement. Medical Care, 41, 1, pp. 130–


