Wilson, Deborah. "Targets and Rankings; Choice and Voice: Performance Indicators and Public Service Accountability." *Public Services: A New Reform Agenda*. Ed. Henry Kippin, Gerry Stoker and Simon Griffiths. London: Bloomsbury, 2013. 145–158. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781472544872.ch-009>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 10:25 UTC.

Copyright © Simon Griffiths, Henry Kippin and Gerry Stoker 2013. You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

Targets and Rankings; Choice and Voice: Performance Indicators and Public Service Accountability

Deborah Wilson¹

Introduction

Targets, rankings and other forms of ‘managing by numbers’ are now a familiar feature of public service performance management. Performance indicators (PIs) are widely used as part of broader accountability mechanisms aimed at improving public service performance. Quantified, public, performance information plays a central role in both top-down, bureaucratic and user-based, democratic accountability mechanisms. Moreover, such mechanisms often run concurrently, utilizing the same performance data, which makes isolating the impact of any one element of these broader governance regimes somewhat difficult. The aim of this chapter is to review the use of PIs within such forms of governance. I first describe the range of ways in which PIs may be employed, and discuss the particular issues associated with their use in the public sector and the range of responses that may result. Second, I explain the different types of PI and discuss the limitations of each. I then consider the use of PIs within, first, bureaucratic accountability mechanisms such as targets coupled with sanctions and/or rewards and, second, within the democratic accountability mechanisms of choice and voice. While the current UK policy discourse emphasizes transparency coupled with choice and voice to achieve public service performance improvement, I argue that the weight of the evidence suggests bureaucratic accountability has been more effective at leveraging such change and discuss the potential implications of this.

¹ CMPO, University of Bristol. This chapter partly draws on previous joint work with Ruth Dixon, Christopher Hood and Carol Propper. Thanks to Lauren Cumming and Ruth Dixon for comments on earlier drafts.

Performance indicators: 'Managing by numbers'

Targets, benchmarks and league tables have a long history in both the private and public sectors. While such metrics are now a familiar feature of public service management, the principle of measuring performance relative to some published measure goes back at least to Jeremy Bentham's plans for prison management in the 1790s (Bevan and Hood, 2006a). Their use in the UK accelerated when Tony Blair came to office in 1997. His administration introduced more than 300 performance targets across all government departments in 1998, directly linked to budget allocations and translated into many more targets and indicators at lower levels of government (Hood, 2006). Since devolution, however, it is England rather than its neighbours that continues to emphasize the use of PIs in its public services performance regimes.

Notwithstanding their widespread use, there are well-documented problems with using metrics-based performance management regimes in the delivery of public services (Propper and Wilson, 2003; Bevan and Hood, 2006a). There is also a growing body of evidence on the shortcomings of both the PIs themselves and on the responses of individuals and organizations to their use in alternative accountability mechanisms (Wilson and Piebalga, 2008). Conversely, there is still much less evidence on whether or not PIs achieve the aim of improving the quality of public service delivery, and very little discussion regarding the relative costs of achieving any such improvement (Propper and Wilson, 2003). Public service outcomes are complex and multi-dimensional. Measurement of improvements in public service quality is therefore difficult, as is attributing their cause to any one specific element of a more general performance management regime, that may use PIs as part of several, concurrent accountability mechanisms.

Governance systems using quantifiable performance indicators such as targets are described by Bevan and Hood (2006a: 518–19) as a form of control 'in which: (1) desired results are specified in advance in measurable form, (2) some system of monitoring measures performance against that specification, and (3) feedback mechanisms are linked to measured performance'. Both (2) and (3), the system of monitoring and the feedback mechanism, can be designed in various, non-mutually exclusive, ways (Burgess *et al.* 2002; Propper and Wilson, 2003).

The measured performance information may be kept internal to the organization or it may be published. There has been an increasing trend for the publication of performance information. If the information is made public

there is a range of possible feedback mechanisms. The information may be used as part of an implicit, or indirect, incentive scheme, under which the organization gets a financial reward as a result of the response of others to the information. A classic example of this is a 'quasi-market', in which providers of services are rewarded for good performance by getting more contracts. More generally, such implicit incentive schemes relate to 'bottom-up' or 'democratic' accountability mechanisms, in which the primary role of quantifiable performance information is to inform and empower consumers who then lever improvement in services through actions such as choice and/or voice.

Alternatively, published performance information may be used as part of 'top-down' or 'bureaucratic' incentive schemes, which explicitly specify how measured performance, often relative to a pre-specified target or to the performance of other provider organizations, is linked to sanctions and/or rewards. These sanctions/rewards may be financial or non-financial, and can be targeted at the level of the individual, the team, the organization or the policy area. So, for example, hospital managers in England faced dismissal if their hospital performed poorly against waiting times targets; conversely good performers were granted earned autonomy and the freedom to keep certain surpluses (Propper *et al.*, 2008). The Quality and Outcomes Framework introduced in 2004 gave family doctors in the UK direct financial incentives to hit targets in the provision of a wide range of specified primary care treatments (Hood *et al.*, 2009). Alternatively, it may be career concerns or reputation effects, with implications that are played out via the labour market, that create the incentive for effort (Burgess and Ratto, 2003). A key point to note is that public service providers often face elements of a range of democratic and/or bureaucratic accountability mechanisms concurrently (Kane and Staiger, 2002). This raises problems of both attributing the effects of the use of PIs to any one mechanism and of minimizing potential conflicts between the incentives created by the different structures (Wilson, 2009).

There are two, linked, assumptions underlying the theory of governance by numbers (Bevan and Hood, 2006a). First, measurement problems are not important: the inevitably incomplete measure of organizational performance contained within a quantifiable performance indicator (PI) adequately represents overall performance. Second, while it changes organizational and/or individual behaviour, such a governance system is not vulnerable to gaming. The particular features associated with the delivery of public services mean that neither assumption is likely to be valid (Dixit, 2002).

The multiple goals and often complex, multi-dimensional, outputs of public service organizations mean that performance relative to these goals is difficult to measure; any quantifiable representation of that output is therefore necessarily incomplete. This creates the opportunity for gaming, the extent of which may depend on individual motivation. Individuals will respond to PIs in ways that maximize their own utility or benefit. This is not necessarily consistent with PIs improving welfare, nor is it necessarily in ways that are expected by those who design the system. In Le Grand's (2003) terminology, responses will differ depending on whether providers are 'knights' or 'knaves.' Moreover, provider motivation may be endogenous to the type of performance management scheme imposed; differently motivated individuals may be attracted to alternative cultures of performance management (Gregg *et al.*, 2008). This relates to the literature on whether public service workers exhibit 'intrinsic' motivation (Frey, 2000; Crewson, 1997; Deci, 1971) and how that may affect their responses to different forms of governance structure.

So neither assumption is likely to hold; managing by numbers in the public services will suffer from both measurement error and the potential for undesired as well as desired responses. The precise way in which these combine with the incentives created by the accountability mechanism(s) employed and individuals' unobserved motivation to create the overall outcomes may in practice be difficult to disentangle (Bevan and Hood, 2006a). Before reviewing the evidence on the effects of using PIs as part of alternative accountability mechanisms, therefore, I first discuss the different forms that PIs may take and the associated incomplete measurement issues that arise.

Forms of performance indicator

The simplest form of PI are raw outcomes, which measure the outcomes of an organization or programme at some designated date – the number of individuals who do not die after emergency admissions for heart attacks, for example, or the number of pupils passing examinations at a certain grade. While they are easy to understand and relatively low cost to collect, they only deal with one dimension of a complex output, and do not provide sufficient information to isolate the impact of the organization on the measured outcome. Patients may have recovered from heart attacks without medical intervention, for example. They are also susceptible to gaming – by adjusting the quality of the intake, an organization can boost its performance as measured by raw

outcomes. Such cream skimming practices are well documented in health and in education and are discussed below. Raw outcomes PIs may unfairly penalize effective providers serving disadvantaged and high cost populations while at the same time hiding poor performance from those serving lower cost populations. Risk adjustment of such measures helps account for heterogeneity in the populations served and therefore helps to both give a better measure of the impact of the agency and reduce the incentive for cream skimming (Propper and Wilson 2003, 2006).

Further to such risk adjustment, the aim of value-added PIs is to better isolate the impact of, for example, the school environment on pupil progress between two points in time. Such PIs do this by incorporating prior attainment and/or other factors which are outside the school's control but which are known to impact on test score (i.e. raw) outcomes. While both reducing the incentive to cream-skim by explicitly accounting for input, and providing a better measure of the effectiveness of an organization, value-added PIs still generally only reflect one dimension of an organization's output; in the case of schools, the focus is still on test scores (or progress between tests at different stages of schooling).

One response to this aspect of measurement error has been the development and increased use of composite indicators that attempt to combine many dimensions of an organization's output into a single figure or rating. Examples include the star rating systems for hospitals in England and the CPA (comprehensive performance assessment) system of rating local government. While intuitively appealing and easy to understand, they are in practice complex and opaque (Hood, 2006) and create various sources of measurement error (Jacobs and Goddard 2007; Stevens *et al.*, 2006; McLean *et al.*, 2007).

Two further points are worth emphasizing here. First, different aspects of a public service organization's performance will matter more or less to its different stakeholders. Both the designers and final users care about a range of outcomes, but the weights these two stakeholders attach to these outcomes may differ (Propper and Wilson, 2003). This has implications for the aggregation methods employed for calculating composite indicators. Or consider a value-added measure of school performance: if it is *not* corrected for school resources (as was the case with the English contextual value added (CVA) measure) it is more suited to informing parental choice than to the government aim of raising standards for the same resource base (raising efficiency). The same PI may not be able to satisfy the often multiple purposes for which it is employed or the different concerns of multiple stakeholders.

Second, we need to distinguish the shortcomings of the PIs themselves from the shortcomings of the uses to which they are put. For example, for PIs to be used as part of meaningful ranking exercises they need to discriminate clearly between different units. While the CVA measure for English schools contained useful performance management information, for example, it was misleading to use it to rank individual schools. Once the uncertainties associated with its calculation are taken into account, Wilson and Piebalga (2008) found that over half of all English secondary schools are not significantly different from the national average when 'ranked' on their CVA scores. Jacobs *et al.* (2007) obtained similar findings for hospital star ratings.

PIs and bureaucratic accountability: Targets and rankings

There is now a large and growing body of evidence on how individuals respond to PIs as part of bureaucratic accountability mechanisms such as targets and rankings. Much of this evidence relates to gaming responses and other unintended consequences of their use, while there is less evidence on whether their use improves public service outcomes. Hood (2006, 2007) discusses the different forms of strategic or gaming behaviour that result from target and ranking systems. Target systems produce the potential for ratchet effects, threshold effects and output distortions. Ratchet effects arise in systems where the target is set as an incremental advance to current output, which creates the incentive to suppress that current output in order to reduce future targets. Threshold effects occur due to the incentive to just hit the target and do no more; and output distortions describe the incentive to focus on hitting the target to the detriment of effort directed at real improvements in performance.

Ranking systems (league tables) may also create similar incentives for output distortion. If the measured performance of an organization is being directly – and publicly – compared with that of others, the incentive is to focus effort on that measured element in order to boost the ranking position. The extent to which this deflects from effort towards true improvement will depend on the extent to which the PI accurately reflects overall performance. The multi-product, or multi-task, nature of many public service organizations' output gives agents opportunities to divert activity away from non-incentivized tasks, which furthers the potential for output distortions (Propper *et al.*, 2008b).

There is a large body of evidence on such undesired responses to league tables and targets in education and health care.² In the education sector, this includes output distortions such as teaching to the test and a narrowing of the curriculum in order to boost published test score outcomes (Wiggins and Tymms, 2002; Jacob, 2002); various forms of cream skimming to boost performance as measured by raw outcomes PIs (Meyer, 1997; Figlio and Getzler, 2002; Cullen and Reback 2002; West and Pennell, 2000); and a focus on 'borderline' pupils in order to meet targets, to the detriment of the outcomes of their non-borderline peers (Deere and Strayer, 2001; Reback, 2008; Burgess *et al.*, 2005; Wilson *et al.*, 2006).

There are also numerous examples of strategic response to targets in the health care sector. Green and Wintfeld (1995), for example, argued that 41 per cent of New York State's reduction in risk-adjusted mortality could be accounted for by data gaming. For the UK, Bevan and Hood (2006a) reported evidence on five types of output-distorting responses to Accident and Emergency waiting time targets. These include requiring patients to wait in ambulances until they were guaranteed to be seen within the four-hour target, and removing the wheels from trolleys to turn them into 'beds' to satisfy the target that patients must be admitted to a hospital bed within 12 hours of admission. Bevan and Hamblin (2009) additionally provide evidence of gaming of the eight-minute response time target by ambulance trusts.

There is some evidence that the use of PIs as part of bureaucratic accountability mechanisms has improved health care outcomes. The most reliable US evidence comes from studies from the New York Cardiac Surgery Reporting System. Hannan *et al.* (1994) found that mortality declined significantly following publication of data on mortality rates, with no accompanying evidence of cream skimming. For the UK, the incentivized targets regime of the Quality and Outcomes Framework appears to have had a positive impact on specific outputs in the primary sector (Hood *et al.*, 2009) although there is additionally evidence of output distortions resulting from this performance regime (Checkland *et al.*, 2008). Propper *et al.* (2008, 2008b) provide evidence that the 'targets and terror' regime on hospital waiting times for in-patient care in England led to shorter waits on average (relative to Scotland, where there was no such incentivized regime), and that this was not at the expense of other, less well-monitored, aspects of patient care. Alvarez-Rosete *et al.* (2005), Bevan and Hood (2006b) and Hauck and Street (2007) also compared waiting times across the devolved

² I review this evidence in more detail in Wilson (2010).

administrations of the UK and found that English waiting time targets had a positive impact.

In education, again there is some evidence of improved outcomes, albeit alongside the evidence on less desired responses outlined above. For the US, the picture is nuanced by student type and by subject. Ladd (1999), for example, found that the Dallas test-score-based accountability programme had a positive effect on outcomes for White and Hispanic seventh-graders, but not for their Black peers. Dee and Jacob (2009) found positive effects on maths scores but no impact on reading scores. Conversely, Jacob (2002) showed that test scores in maths and reading both increased in Chicago public schools after the introduction of a test-based accountability policy, but that student effort and improvements in test-specific skills largely drove those improvements. Hanushek and Raymond (2005) found a positive effect of 'consequential' accountability, whereby published school performance was linked to a range of consequences, including vouchers for pupils to leave poorly performing schools and the threat for a school of potential takeover.

Similarly in the UK, school performance as measured by publicly available PIs has potential consequences via pupil choice and more top-down sanctions such as potential takeover. Burgess *et al.* (2010) provide evidence of the effectiveness of school league tables using this UK data. From 1992 to 2001 secondary school league tables were published annually in England and Wales. In 2001 the Welsh Assembly Government abolished league tables, thereby removing a key element for accountability of two otherwise near-identical education systems. By comparing student outcomes in England and Wales before and after 2001 the authors were able to isolate the impact of that specific policy change. They found that abolishing school league tables markedly reduced school effectiveness in Wales relative to England, by an average of almost two exam (GCSE) grades per student per year. This effect was concentrated in the schools in the lower 75 per cent of the distribution of ability and quality, with schools in the top quartile showing no effect. The policy reform in Wales therefore reduced average performance and raised educational inequality relative to England. Given the largely rural geography of Wales, and the subsequent practical limits on the potential for parental choice, the authors concluded that it was the use of league tables within the bureaucratic accountability framework, rather than the concurrent democratic accountability, that was the most likely driver of the results. This is despite the current focus of policy discourse being centred on the use – and potential effectiveness – of performance information as part of democratic accountability frameworks incorporating choice and/or voice.

PIs and democratic accountability: Choice and voice

Since the 1980s user-based mechanisms such as choice, voice and personalization have been introduced by governments as part of successive public service reforms, both in the UK and elsewhere. Such initiatives generally have the aim of improving public service quality through consumer pressure, as well as empowering users, making public services more responsive to their needs. Choice and voice are considered to be complementary: 'Providers will be held to account through a combination of mutually reinforcing choice, voice and transparency mechanisms, depending on the service being provided' (Cabinet Office, 2011: para 5.26). This discourse about choice and voice working together to improve quality can be traced back to Hirschman (1970), who argued that exit (choice) and voice are two consumer responses to deterioration in the quality of a firm's product or service. Exit is 'the sort of mechanism economics thrives on' (Hirschman, 1970: 15). It is neat, impersonal and indirect: subsequent recovery by the firm comes via the market. Voice, by contrast, is more 'messy', more personal and more direct, and for Hirschman can cover anything from personal complaint to collective action. So, exit belongs in the economic realm, with consumer decisions in the marketplace as levers of change, while voice is positioned in political, democratic realm with individuals as citizens making their feelings known (Dowding and John, 2008; Greener, 2007).

Hirschman argues that a process of decline in the quality of a firm's output (for whatever reason) activates certain consumer responses – exit and voice – which in turn act as endogenous forces of recovery, thereby reversing the initial decline in quality. This is a self-correction mechanism, whereby the very process of decline activates certain counterforces and hence generates its own cure – the 'embedded incentives' in Le Grand's (2007) terminology. Hirschman argues that different combinations of exit and/or voice are suitable in different settings, depending on which signal(s) the firm is most responsive to.

In practice, despite the simple appeal of 'choice' and 'voice', the terms are actually used in many different ways and can refer to quite different institutional arrangements, which in turn impact on the outcomes of any one user-based mechanism. Three questions serve to frame the basic elements of a choice-based system. First, who chooses? In education it is the consumer herself, the pupil, or the parents acting on her behalf. In the health care context the choice is more usually delegated to a specialized agent (gatekeeper, intermediary, payer) acting on the consumer's behalf. In practice, in health care, patient choice is

often somewhat limited, with payer choice dominant, although the distinction is sometimes blurred (Propper *et al.*, 2006).

Second, what do they choose? In state-funded systems, price is generally not the determining factor and quality is important. Parents are likely to care about the quality of education, broadly defined, which may include factors such as a school's previous results, its composition, ethos, facilities, location (Wilson, 2009; Burgess *et al.*, 2009, 2011; Allen and Burgess, 2010). In health care, the patients are often fully insured against the price and so choose on the basis of quality, while the body responsible for buying care will be interested in both. Finally, what are the constraints on the process of choice? For parents the most obvious constraint is that their preferred school is full. Transport time and costs may be a constraint on choosing other schools, and house prices may be a constraint on moving. In health care, travel costs and other costs of using care may constrain choice. Consumers may feel 'locked-in' to local government services, with little realistic option for exit. In all public service contexts the lack of accurate information on the quality of the often complex service being provided may limit the extent to which consumer choice, or voice, creates the incentive to improve that quality.

While user empowerment through choice and voice may be desirable in and of itself, it is hard to imagine that users will continue to be satisfied with having some form of control in their public services if there are no noticeable improvements in their experiences of them. This relates to the distinction made by Dowding and John (2009a) between intrinsic value – choice being desired for itself; and instrumental value – if it results in welfare gains in the form of productive and/or allocative efficiency. They argue that attention should be concentrated on the potential welfare gains of choice rather than any supposed intrinsic benefits.

The use of choice and, increasingly, voice as a means to achieve such welfare gains is now well established in the UK and elsewhere (Burgess *et al.*, 2007; Propper *et al.*, 2006; Simmons *et al.*, 2009). Le Grand (2007) argues that such mechanisms create a system with incentives for reform embedded within it. Choice, coupled with a system of funding in which resources follow the user, applies competitive pressure to providers who therefore have the incentive to improve their service in order to attract or retain clients. Such improvements may include increased responsiveness to the needs of individual clients, which is often discussed as part of the 'personalization' agenda. Another goal of user-based mechanisms has been to improve equality of access for those users who historically have found it harder to access good quality services. Le Grand

(2007) goes further and argues that choice is better than voice at achieving this. So, for example, school choice can, in theory, break the link between house prices and access to 'good' schools by reducing the importance of geographical proximity as an allocation criterion. In practice this has not been so successful (Burgess *et al.*, 2009, 2011).

Whether we think of user-based mechanisms solely as means of empowering users, or as a means by which the quality of public services are improved, a fairly obvious consideration is whether people want choice and, if so (how) do they respond? Curtice and Health (2009) and the 2020 Public Services Trust (2010) both found that people want choice, but that it is not necessarily a priority. They found less support for diversity of provision, with only a minority of respondents in favour of private companies running either (state funded) hospitals or schools.

Propper *et al.* (2006) reviewed the evidence on whether patients responded to greater choice in health care (see also Dowding and John, 2009a).³ They concluded that direct patient choice is limited in many systems, and that it may conflict with choice exercised by the agents who place contracts with hospitals on behalf of groups of patients. Patients in England have expressed willingness to travel to non-local hospitals and have done so when given assistance to exercise this choice. When such support is absent (or the wait at the local hospital is perhaps shorter or less uncertain), the evidence from European countries suggests there is relatively little take-up of such travel options. Individuals who are better informed and individuals whose illnesses are more severe may be more likely to travel and this may lead to greater differences across hospitals in patient severity.

This points to differences across individuals in their exercise of choice. There are also potential issues regarding differences in individuals' basis for choice. While the theory behind school choice, for example, rests on parents choosing at least partly on the basis of (published) academic standards, which in turn creates the incentive for schools to improve those standards, it may be that parents (additionally) have what 6 has termed 'segregationist preferences' (6, 2003 and references therein; Allen and Burgess, 2010), whereby individuals choose schools (or other public services such as housing) on the basis of trying to avoid certain types of other user. Recent evidence from Burgess *et al.* (2009, 2011) shows that parents choosing a primary school in England do value academic standards, but they also choose on the basis of proximity and

³ I review the evidence on choice in health and education in more detail in Wilson (2010).

a school's socio-economic composition, preferring schools with lower proportions of pupils from low income households. I return to the issue of choice across multiple dimensions of quality below.

Does choice-based accountability improve outcomes? There is some evidence from the UK reforms of the 2000s that quality improved in hospitals located in more competitive areas, and that such improvements may have been due to better management in those hospitals (see Propper, 2012 for a review). However, the extent to which these results can be generalized across different aspects of health care has been questioned (Bevan and Skellern, 2011) and the evidence suggests that, while consumers claim to want information on hospital performance, they do not in practice make great use of the data. In education, it is far from clear that school choice has improved the academic performance of schools (Allen and Burgess, 2010). It is difficult to disentangle the extent to which this is a failure of choice *per se*, or a result of the constraints on the ways in which choice has been realized in practice.

While there is now a substantial body of evidence on choice in public services there is much less on voice, or on how choice and voice may work together (Dowding and John, 2008, 2009b). Wilson (2009) applies the original Hirschman framework to the education context, and shows that it is only under restrictive assumptions regarding how 'quality' is measured that choice and voice complement one another to improve that quality for all. If, as is more realistic, school quality is considered to comprise many dimensions this may no longer be the case. Wilson outlines the case where multi-dimensional school quality coupled with league tables focusing on raw outcomes PIs creates the incentive for schools to respond to either choice or voice signals from only one type of (middle-class) user. More generally, her analysis emphasizes how concurrent accountability mechanisms employing an imperfect measure of quality or performance may create potentially undesirable responses from provider organizations.

Discussion

So, what have we learnt from the use of PIs in public service delivery to date? The evidence suggests that PIs are most effective within a system of bureaucratic accountability. Providers do respond to the public disclosure of performance information when it is part of an incentive structure comprising financial and/or non-financial rewards and sanctions. There is some evidence that targets and

such ‘carrots and sticks’ work, particularly if the desired outcome is focused and measurable, as in the case of hospital waiting times. There is less evidence that democratic accountability via choice and/or voice works to systematically improve outcomes for all users. In education, for example, the evidence suggests that the mechanism via which the league tables improve school performance is one of reputation rather than parental choice (Burgess *et al.*, 2010). A common theme that emerges is the extent to which the outcomes of any one accountability mechanism are situation specific, which suggests caution is required before translating a policy from one institutional context to another.

The two assumptions underlying PI-based governance structures do not hold for public service delivery: measurement error is an inherent problem, as is the resultant potential for undesired as well as desired responses, and the evidence bears this out. Using such PIs within either a bureaucratic or democratic accountability mechanism provides the incentives for both kinds of response. How that is manifested will depend on the form of PI that is employed, the use to which it is put (in terms of targets set, rankings created), and the unobserved motivations of provider organizations and employees. The current UK policy discourse emphasizes increased transparency via publicly available performance information, coupled with choice and voice as drivers of improved standards for public service delivery (Wilson, 2012). Given that the evidence suggests that PIs are most effective within a system of bureaucratic accountability, the danger is that providers may focus their efforts on attempting to rank highly on potentially conflicting performance measures to avoid bureaucratic sanctions, rather than responding to the needs of the full range of individual service users.