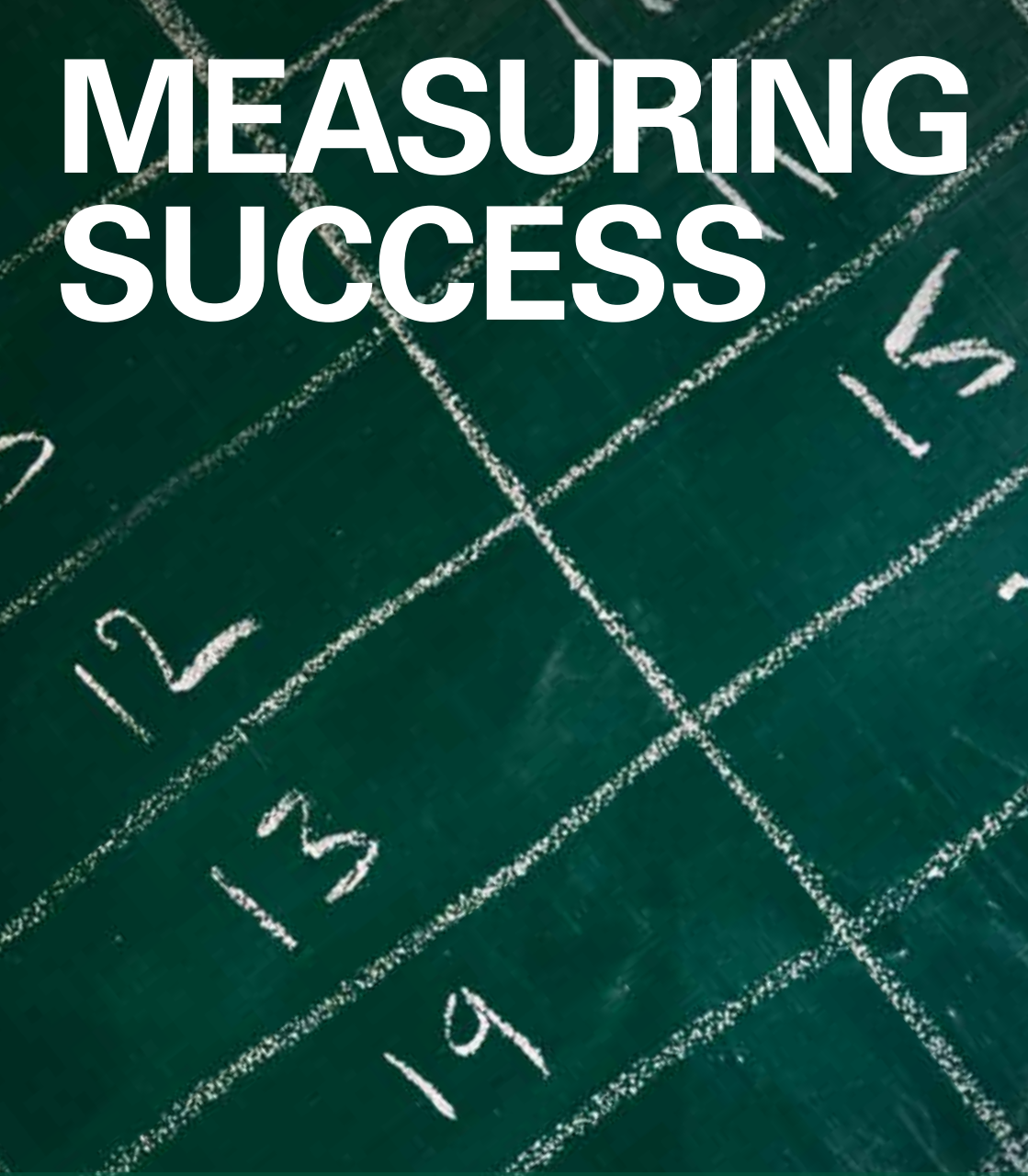

MEASURING SUCCESS

League tables in the public sector

Beth Foley
and Harvey Goldstein

BRITISH ACADEMY
POLICY CENTRE

Measuring Success

League tables in the public sector

Beth Foley
and Harvey Goldstein
March 2012

THE BRITISH ACADEMY
10–11 Carlton House Terrace
London SW1Y 5AH
www.britac.ac.uk

Registered Charity: Number 233176

Published March 2012

ISBN 978-0-85672-600-2

Designed by Soapbox, www.soapbox.co.uk
Printed by Smith & Watts

Contents

Acknowledgements

We would like to thank Stephen Finnigan, Rob Copeland, Mick Brookes, Ann Mroz and Aaron Porter who were interviewed in the early stages of the project. We would also like to thank all those that participated in a policy forum held in January 2011 to feed in to the project (a list of attendees is included at the end of the report), the members of the steering group: Professor Stephen Ball FBA; Professor Colin Crouch FBA; and Professor David Bartholomew FBA, five anonymous peer reviewers and Helen Haggart at the British Academy for their comments and feedback on earlier drafts.

About the authors

Professor Harvey Goldstein FBA is Professor of Social Statistics in the Centre for Multilevel Modelling, Graduate School of Education, at the University of Bristol. Professor Goldstein is a chartered statistician, has been editor of the Royal Statistical Society's Journal, *Series A*, a member of the Society's Council and was awarded the Society's Guy medal in silver in 1998. He was elected a member of the International Statistical Institute in 1987, and a fellow of the British Academy in 1996. He was awarded an honorary doctorate by the Open University in 2001.

Beth Foley is a researcher at the Institute for Employment Studies. Prior to this, she spent two years as a Researcher for the Social Market Foundation think tank, carrying out projects on public service reform, mental health and welfare reform, behavioural economics and its impact on policy-making, and the issue of insecure employment. She has also worked as a freelance researcher for both the British Academy Policy Centre and for the Local Authorities' Research Councils Initiative.

Executive summary

Why league tables?

Over the past three decades in the UK, demand has grown for accountability and user choice in relation to public sector institutions. Growing out of the performance management movement in the private sector, and aided by the increasing availability of large administrative databases, the most visible manifestation of this has been the publication of institutional rankings or 'league tables' based upon particular performance indicators. League tables are now widely used in the public sector, and have been employed in health, social services, policing and education. Given the variety of public sector institutions in which performance indicators are now employed, it was not possible to cover all of these areas in this report, so it is the last two of these – policing and education – that are discussed here, and where the British Academy, among its Fellows, has considerable expertise.

Given their ubiquity and increasing importance, an account of the provenance, the strengths and the weaknesses of league tables is overdue. A fundamental problem that surrounds discussions of public sector performance monitoring is the lack of systematic evaluations of its effects. This absence of sound evidence has made performance measures a highly contentious area, where different viewpoints have developed. This report explores the issues raised by public sector performance monitoring, to provide the basis for a more informed debate about its use and to ensure it best serves the policymakers, professionals and public service users of the future.

Performance rankings are intended to serve two purposes. The first can be described as 'public accountability', whereby those who provide resources to run institutions such as schools or police forces can form judgments about where improvement is needed or particular action is required. The second is to provide users of services, such as parents

who wish to choose a school, with information to assist them. In both cases it is envisaged that institutions themselves, as well as external agencies, will react to published rankings in ways that enhance performance by encouraging competition between institutions in a quasi-market environment. Proponents often point to examples such as the improvement in examination grades following the introduction of school league tables as evidence for their positive effects.

A third, and not so obvious, function of league tables and their associated 'institutional targets' is that of control. Providing targets, such as those associated with school examination results, is seen as a powerful means of making policy indirectly by providing appropriate incentives for behavioural change and the report describes some examples where this has happened.

Supporters of league tables also appeal to democratic openness, suggesting that giving citizens good access to statistical information will lead to greater participation in decision-making, and that access to public data should be a democratic right. The opening up of government databases generally, as well as the provision of league tables, is seen as part of this movement.

Critics of league tables have several reservations. The first is that, while agreeing that publication will tend to change behaviour, they argue that this is often associated with perverse 'side effects' that are deleterious, and that important areas may be ignored following excessive focus on improving league table positions. Thus, for example, concentrating on a reduction of headline figures for particular types of reported crime may lead to excessive neglect of other areas of policing by removing resources from them. There is evidence that schools engage in 'gaming' to improve their ranking, by manipulating exam entry policy to the detriment of student choice, or even by excluding low achievers. Secondly, critics suggest that the range of what is measurable and hence amenable for use in performance indicators is limited, and concentration on these detracts from other, less quantifiable objectives such as breadth of learning. They also point to what they see as the arbitrary way in which, for example at university level, individual indicators are aggregated to produce a single 'one-dimensional' ranking.

Finally, critics suggest that there are two major technical issues that substantially weaken the case for publication of rankings. The first is that any ranking needs to be contextualised. Thus, higher education rankings

of degree results should be adjusted for differential selection of students and school examination results need to be adjusted for the intake achievements of students when they start at a school – so called 'value added' rankings. The second issue is that the uncertainty surrounding any given ranking is very large, and in many important cases so large that no statistically meaningful comparisons can be made, nor can useful user choices be sustained.

Scope of the report

The report looks at league tables for schools in rather more detail than league tables for other areas, because these are the most developed and have the longest history. School league tables provide the most extensive data for researchers to study, enabling them to examine their statistical limitations and to quantify the uncertainty that critics suggest makes their use for both accountability and user choice highly problematic. The study of school league tables is important in seeking ways to enhance the positive aspects of performance monitoring and to improve its use.

The more recent introduction of league tables in higher education clearly illustrates how parts of the media have taken initiatives to compile league tables and how this has become a global activity. In contrast, it also provides an interesting example of how an independent institution can manage such tables with a large degree of integrity in terms of presenting both advantages and reservations.

In the area of policing, the issue of what is being measured is particularly apposite since there is an acknowledged diverse set of criteria. This area reveals some examples of how a government department is able to engage with the issues of improvement through feedback of information via an inspectorate rather than public rankings – an example of what we might label as 'intelligent accountability'.

In all the areas discussed detailed consideration is given to technical issues in a manner that ensures they are accessible to non-professionals.

While the scope of this report is largely limited to the UK and the recommendations aimed at a UK audience, it does draw on international experience. Much of the discussion will be of interest and relevance for a number of countries.

Conclusions

League tables certainly affect behaviour. In some cases this may be for the good, but not universally. The government, which has been largely responsible for promoting these tables, must think more carefully about their use and give attention to raising public awareness of their properties. There is a need to evaluate their functioning in a broad sense so that their best aspects can be preserved while limiting their drawbacks. It is also important to stress the need to address the statistical limitations of league tables: if they are statistically unreliable this will inevitably undermine whatever strengths they may have. If their use is to be continued, some of the issues surrounding league tables outlined in this report will need to be thought through and addressed, to ensure they meet their aims and best serve policymakers, professionals and the general public.

Recommendations

The report points out that good evidence about league tables is in short supply. Evaluations of existing uses are rare, as are pilot studies before full implementation. This has resulted in the use of anecdotal evidence, much of which is critical, pointing to perverse side effects, 'gaming' and the like.

However, some pilot studies and international examples provide important lessons. It is in this context that the report has the following recommendations:

General

- Serious consideration should be given to using comparative rankings as 'screening' devices that are not published or made available beyond those institutions involved, but used as part of an institutional improvement programme, so that institutions can seek improvement without perverse incentives arising from full public exposure. We refer to this as 'intelligent accountability'. This could obviate some of the currently perceived negative effects of league tables.
- Wherever league tables are published they should be accompanied with appropriate and prominent 'health warnings' highlighting their technical limitations. These should include assessments of the statistical uncertainty, often large, that may limit their usefulness. They should also include statements about the quality of the meas-

urements that go to make up the indicators, including the effects of aggregation. In a broader context, there is a need for a debate about whether simply making data available to citizens will encourage good use of them. In the absence of professional support and advice, data analysis can be very difficult for those with limited experience or expertise. Deliberate or unintentional misuse of statistical information should not be encouraged and there is a real danger that this could occur increasingly unless public awareness of the issues improves.

- More research is needed on the effects of performance data on institutional performance. There should be careful evaluation of existing league table systems and the systematic piloting of proposed systems. This evidence should pay particular attention to 'knock-on' effects whereby resources may be reduced for some important activities in order to improve league table performance.
- Consideration should be given to whether one or more independent (not for profit) institutions could have a role in monitoring developments, providing guidelines for good practice and also become involved in the production and presentation of performance indicators. Such institutions should be independent of government.

Education

- The linking of league tables to rewards should be weakened to reduce the side effects of inappropriate 'gaming' and to reduce stress among teachers, parents and students. This would also have the desirable effect of making the results a more objective evaluation of performance. The problematic consequences for schools serving the most disadvantaged pupils particularly need to be addressed.
- The government should consider ways to prevent league tables being exploited by the media, such as ensuring that measures of uncertainty are provided around any institutional results. Associated with this there could be a campaign to better inform the public at large about the strengths and limitations of league tables, although any such attempt poses considerable challenges.
- Consideration should be given to alternative ways of using quantitative information to monitor educational performance generally. This can be achieved by in-depth study of a sample of schools and students within a national database. A useful model is the Assessment of Performance Unit that was set up in the 1970s in England and discontinued in the 1980s (Gipps and Goldstein, 1983).
- Consideration should be given to using performance information as a screening device rather than publishing as league tables, as in the

Hampshire experiment. This could be accompanied by an emphasis on evaluation and inspection systems that are designed to emphasise ways of assisting schools to cope with problems rather than 'exposing' them using public rankings.

- Ways to rely less on a small number of indicators should be sought, as well as those which cover more aspects of learning.
- More appropriate statistical analysis models should be used to describe institutional differences that allow for differential performance for different groups of students. In particular, there should be a shift away from the comparison of individual institutions towards research that helps to identify modifiable factors that appear to be related to good performance.
- An ethical code to govern the publication of school performance measures should be formulated, as suggested by Goldstein and Myers (1996). This would be based on two broad principles: that unjustified harm to those to whom the information applies should be prevented, and that there should be no *absolute* publication rights for performance data.
- Further consideration needs to be given to the role of inspection and accreditation agencies as a means of evaluating individual institutions. Trust in such agencies may not be easy to achieve, especially when they are perceived to be instruments of government. A discussion of such agencies is given in Appendix A.

Higher education

- Indicators need to be selected according to validity rather than availability as currently tends to be the case. This implies more qualitative and process indicators, although care needs to be exercised in terms of their subjectivity.
- Disaggregated indicators are important and the temptation to aggregate into one index, or even a small number of indexes, should be resisted.
- Measures of uncertainty need to be displayed.
- For users, broad categories rather than precise rankings are to be preferred and sensitivity analyses with different weightings to components should be conducted to test stability.
- Subject-based rankings should be emphasised.
- Further consideration needs to be given to the role of inspection and accreditation agencies as a means of evaluating individual institutions. Trust in such agencies may not be easy to achieve, especially when they are perceived to be instruments of government. A discussion of such agencies is given in Appendix A.

Policing

- Current indicator measures on crime do not fully account for the heterogeneity of policing environments and challenges within different force areas. Crime should be contextualised in terms of local conditions.
- Problems with the recording of crime, especially where high stakes targets are in place, need to be addressed.
- The uncertainty attached to statistical estimates, especially for small areas, should be addressed.
- Increased accountability to the government is perceived to be encouraging a more 'reactive' policing style that is less engaged with the local community. Locally defined outcomes should be incorporated.
- It is only recently that the accessibility of police performance data to the wider public has begun to be explored. There should be better information for public understanding of the complex variety of data.
- As in education, one of the major concerns is the potential for perverse incentives or behaviour in a police force dominated by performance monitoring. There should be thorough evaluation of side effects and perverse incentives.
- There is some evidence that the perceived 'neo-liberal culture' imposed by public sector performance monitoring is also a matter of concern to police officers themselves. The relevance of the idea of 'competition' in a market sense among police forces should be examined.
- The role of unpublished rankings available to the inspectorate for discussion with individual police forces should be explored.

Introduction

In 2009, Conservative Leader David Cameron made a speech on his party's plans for expanding political accountability – what he termed, 'giving power back to the people'. In it, he revealed a central part of this agenda would be 'setting data free':

> *'In Britain today, there are over 100,000 public bodies producing a huge amount of information … Most of this information is kept locked up by the state. And what is published is mostly released in formats that mean the information can't be searched or used with other applications, like online maps. This stands in the way of accountability … We're going to set this data free. In the first year of the next Conservative Government, we will find the most useful information in 20 different areas ranging from information about the NHS to information about schools and road traffic and publish it so people can use it. This information will be published proactively and regularly – and in a standardised format so that it can be 'mashed up' and interacted with. What's more, because there is no complete list that can tell us exactly what data the government collects, we will create a new 'right to data' so that further datasets can be requested by the public. By harnessing the wisdom of the crowd, we can find out what information individuals think will be important in holding the state to account.'*
> (Cameron, 2009)

Since the election of the Coalition Government in May 2010, there have indeed been commitments to a large expansion of the amount of data to be made available, much of it online. This includes data from central and local government and covers all activities, including financial ones (http://data.gov.uk/). However, the collection of comparative data for political and economic purposes is not a new development. Historically, comparative rankings have included economic performance indicators and, in nineteenth century England and Wales, comparisons among primary schools on the basis of the performance of their pupils.

Since the 1980s, however, the rise of doctrines such as New Public Management and the 'reinventing government' movement of Osborne and Gaebler (1992) have led to a rapid growth of interest in the measurement of performance of institutions in the *public sector*, driven partly by the ideology of a competitive marketplace for consumers.

The idea that the quality and efficiency of public service provision could be improved via market mechanisms has resulted in a drive to collect quantifiable information on public sector performance. With the development of large administrative databases, this task has become more straightforward. The political focus on the expansion of 'choice' for public service users has also made the argument for external evaluation more persuasive. In the UK, data is now collected across the public sector, from schools and hospitals to police forces and prisons. In the case of higher education, institutional rankings also include an international dimension.

Public sector performance monitoring now encompasses a variety of quantitative and qualitative measurements. In qualitative terms, we have seen the rise of independent audit and inspection bodies, such as Ofsted and Her Majesty's Inspectorate of Constabulary (HMIC). Their impact has, in turn, generated extensive analysis, but it will be beyond the scope of this report to cover all of these debates here. This study will, instead, focus more closely on the *quantitative* aspects of performance monitoring in the UK. These include methods which rely on the compilation and assessment of performance data, such as star ratings, organisational report cards, targets and league tables.

The choice of method in performance monitoring depends largely on the *type of improvement* that is being sought (Hood, 2007). **Targets** measure performance against a specific threshold standard. Typically, targets are used to select a group of institutions for special attention, for example to allocate resources or to intervene in specific ways. They are generally selected in cases where the focus is on the achievement of baseline standards, and can be an effective and direct way of meeting that particular goal. One example is their use to cut waiting times for hospital treatment in England (Propper *et al.*, 2010). While important, the use of indicators to define targets will not be discussed extensively, although some examples will be given.

Ranking systems, or **league tables**, evaluate the performance of a unit or institution against that of other comparable units. Roberts and Thompson (2007) define a league table as *'*a published set of quantitative data

designed to present comparative evidence regarding the quality and/or performance of organisations'. Ideally, the most significant measures of quality and performance are established, and data reflecting these standards (indicators) are then compiled. Visscher (2001) defines a performance indicator as 'a number of means by which the quality of the functioning of an institution or a system (e.g. a policy area) is expressed'. Individual indicators may be combined for the purposes of ranking institutions. Rankings tend to be deployed to put pressure on service providers to improve their overall performance without specifying particular baselines.

"A fundamental problem that surrounds discussions of public sector performance monitoring is the relative lack of systematic evaluations of its effects and whether its stated aims have been achieved"

Finally, Hood (2007) also introduces the idea of '**intelligence systems**', which gather background information on the quality of performance but do not subject it to a fixed interpretation, unlike targets or rankings. This type of monitoring is likely to be employed where the intention is to improve knowledge about the factors affecting the performance of a system, without focusing on particular measures or incentives to affect the behaviour of the actors in that system.

The results of these different performance monitoring initiatives are varied and have not always been well understood by policymakers. In fact, a fundamental problem that surrounds discussions of public sector performance monitoring is the relative lack of systematic evaluations of its effects and whether its stated aims have been achieved.

In this report we will consider the technical aspects involved in measuring and presenting performance data, how accessible such presentations are currently, and also the implications of publishing comparative rankings for individuals, institutions and society generally. Given the variety of public sector institutions in which performance indicators are now employed, the review will focus on the evidence regarding education – both for schools and higher education – and on policing. These are the two important areas covered by the expertise that resides within the British Academy's Fellowship. Many of the issues are common to other areas such as health, although there are also differences, some of which

we will note. For our two chosen areas we might broadly characterise 'education' as league table-dominated performance culture and 'policing' as a performance culture more focused on targets. For each of these areas we summarise and comment on the existing evidence issues.

The use of performance measures is a highly contentious area where, because of the lack of sound evidence, quite different viewpoints have been adopted. We examine the negative aspects of performance measures that critics say have emerged as 'side effects' of their publication, and evaluate the importance of these side effects. Based upon experience around the world, the report also suggests ways in which the publication of such data can be enhanced while minimising any undesirable aspects.

The report explores the ongoing technical, political and societal questions raised by public sector performance monitoring in an effort to provide the basis for a more informed debate about its use and to ensure it best serves policymakers, professionals and public service.

Why publish data?

The publication of performance data has three broad aims.

1. A **control function**, allowing governments, public sector managers and the general public to monitor the functioning and, in particular, the efficiency of public services. This enables public discussion, especially debate about whether public money is being used effectively.
2. A **market function**, providing service users with information that can be used to choose between different providers.
3. An **improvement function:** this is intended to change behaviour by encouraging service providers to review and seek to improve their performance. We shall also discuss the use of performance data that remain *unpublished*, as one of the possible alternatives to current league tables.

What are the potential benefits?

Proponents of public service performance monitoring argue that the availability of performance data has the power to alter, and in some cases im-

prove, an organisation's priorities and behaviour. Evidence does suggest that public service professionals respond to targets and league tables; following the introduction of benchmark targets for GCSE performance, the percentage of pupils achieving the standard of five A*–C grades at GCSE has been steadily increasing. In health, Propper *et al.* (2010) found that, in comparisons between patient waiting times in English and Scottish hospitals, those in England – where targets for reducing waiting times had been imposed – performed better than those in Scotland, and that there was little evidence of system 'gaming'. Interviews with public service professionals also revealed that, when presented appropriately, performance data can highlight potential problems and had encouraged them to focus more closely on the quality of the services they provide.

What are the problems?

While performance indicators do appear to have the potential to act as a tool to improve organisational performance, their rise has come in for criticism, both on theoretical and empirical grounds. As Hallgarten (2001) points out, 'it should come as no surprise that performance indicators change an organisation's priorities. That is precisely their purpose. The concern occurs when such indicators skew priorities to the extent that other, normally less measurable, goals are relegated or jettisoned'. Partly because of the lack of appropriate evaluations of their usefulness, the drawbacks of league tables have inevitably tended to be highlighted as opposed to their strengths. There is thus more evidence about their limitations.

Within academia, research has highlighted a number of issues with the measurement of public sector performance. Wilson, Croxson and Atkinson (2006), for example, point out that public sector bureaucrats must serve multiple stakeholders, including service users, taxpayers and politicians. These stakeholders hold diverse and frequently conflicting aims, making overall measures of performance and quality very difficult to define.

This, in turn, makes performance relative to multiple and sometimes vague goals hard to measure. The complicated causal link between measuring and improving public service performance leads Wilson *et al.* to brand performance indicators 'an imperfect measure of a complex process' (2006, 154). As Tilley (1994) points out, 'the term "performance indicator" needs to be taken seriously. An indicator is a pointer. It is not

a direct measure … Indicators are better or worse depending on how precisely they tap the underlying feature they are supposed to assess'. Thus we need to recognise that indicators are just that: they do not constitute a definitive judgment on institutions (or individuals), and we shall return to this issue later.

> "Both ranking- and target-based systems are also vulnerable to 'gaming'"

A consequence of this situation is that, as Propper and Wilson (2003) point out, public sector professionals generally have 'several ends to achieve' (251–2). They will be expected to produce both efficiency and equity. They may also be more risk averse and more motivated by non-pecuniary benefits, based on an altruistic commitment to providing a public service. The risk is that these 'intrinsic' motivations exhibited by many public sector workers can be 'crowded out' through an emphasis on extrinsic rewards.

Both ranking- and target-based systems are also vulnerable to 'gaming', defined by Hood (2007) as the 'deliberate massaging or outright fabrication of numbers collected with the intention of improving the position of an individual or organisation'.

Empirically, there have been a number of questions raised around the utility of performance indicators. Burgess, Propper and Wilson (2002) seek to investigate these implications via a study which involved monitoring the effects of performance management via league tables on store managers within a supermarket chain. Their results revealed that the introduction of league tables quickly transforms league table position into an overriding goal amongst managers. This, in turn, encourages more risk-seeking behaviour; where managers perceive an opportunity to move to the top of the league table, they tend to discount the organisation's general objectives and will even accept proposals which may be detrimental to broader goals. Given the potentially adverse effects of this 'measure fixation', they advise caution in the use of league table position as a driver of performance.

While we need to be careful about generalising such results, this study does raise issues that are relevant to public service provision. Indeed, there are a number of possible unintended consequences of performance monitoring.

Smith (1995) lists eight important problems which performance monitoring may generate:

- **Tunnel vision**: a managerial focus on quantifiable phenomena at the expense of all others. The problem of unquantifiable objectives is particularly acute in the public sector, given the diversity of stakeholder preferences.
- **Sub-optimisation**: the pursuit of narrow local objectives at the expense of the aims of the organisation as a whole.
- **Myopia**: performance indicators provide a snapshot of organisation activities, but ignore long-term developments and consequences. This makes them an imperfect tool for assessing current management practice, due to the cumulative nature of developments.
- **Measure fixation**: a focus on measures of success rather than underlying objectives.
- **Misrepresentation**: deliberate manipulation of the data collected.
- **Misinterpretation**: accidental misreading of the data, or unawareness of its limitations.
- **Gaming**: deliberate manipulation of behaviour to maximise league table position.
- **Ossification**: organisational paralysis due to an excessively rigid system of performance management.

There are also a number of issues in the reporting of performance data. Public sector performance monitoring must navigate three important trade-offs. The first is that between the accessibility and intelligibility of the information and measures used, and the accuracy of that information. The second is between the availability of information and its validity as a performance measure. The third is that between qualitative and quantitative measures. There are also technical questions of adjustment and reliability (Bird *et al.*, 2005, Goldstein and Spiegelhalter, 1996, Leckie and Goldstein, 2009) that limit the inferences that can be legitimately drawn from any ranking, whether for the purpose of institutional accountability or for user choice.

This is not to say that these kinds of issues always occur in practice, and indeed there are a range of other dangers associated with an entirely unregulated system. However, it is vital for policymakers to remain aware of such potential problems within performance monitoring frameworks.

1 Education

In the public sector, school-level education is the field in which league tables have been established longest and exhibit their most pervasive influence. Performance data were seen as a means to facilitate parental choice and to encourage professionals to re-assess the quality of their teaching; as one teacher interviewed put it, the introduction of league tables into schools did initially provide a 'system shock' and a 'push' for teachers to raise their expectations and focus on performance. In recent years higher education has also seen a rapid growth of league tables for institutions and individual disciplines. With the number of universities increasing dramatically and the introduction and rise of tuition fees, the demand for league tables has been couched in terms of a growing 'market' in higher education. This review will set out the characteristics of performance monitoring and league tables in schools and higher education in turn, as well as recommendations to improve their use and to tackle the more problematic aspects of this system.

Schools

The use of league tables to measure the relative performance of schools according to their pupils' achievements in standardised national exams has been one of the most significant and, in some quarters, controversial educational developments of the last 20 years. Since the introduction of national systems of testing and teacher assessment via the National Curriculum, the role of government has been to collect data on schools' test scores and exam results, using examination group data, the Pupil Level Annual School Census (PLASC) and the National Pupil Database (NPD). School and college results are published alphabetically, and these can then be ranked by numerous media outlets.

School league tables have been in existence since 1992. They have also come to be viewed by government as an important aspect of the pro-

cess of policy formation. For example, the 'Parent's Charter', introduced by the Major administration in the early 1990s, explicitly recommends the use of such league tables for school choice. The data has also been used for target-setting: for example, the National Challenge, a scheme announced by the Labour government in 2008, was based explicitly on the key indicator employed by league tables (the percentage of pupils gaining five or more GCSEs at grades A*–C). Schools which fell below the 30% threshold on this measure were singled out for additional evaluation and support.

However, schooling has also been the area in which the purpose and practice of league table construction has come in for the most sustained criticism in recent years. A central problem is the lack of satisfactory objective evaluation of the long-term effects of these tables. Across the primary and secondary sectors, professionals have consistently voiced their concern about the technical limitations of league tables, the misconceptions which continue to exist around their measurements and the adverse effects that the growth of a league table culture is having on educational processes and outcomes. In some educational systems, for example within the UK in Wales, Scotland and Northern Ireland, there has been a drawing back from the publication of such tables (Wiggins and Tymms, 2002; Russell, 2001).

League tables for secondary schools are currently based on three key indicators: 1) pass rates at GCSE; 2) measures of value-added; and 3) absences (authorised and unauthorised). It should be noted that almost all league tables published in the media place heavy emphasis on the first of these indicators. League tables in the primary sector are based upon Key Stage 2 tests taken in year 6. Their use in the primary sector is particularly controversial. Scotland has never published primary school league tables and they have been scrapped in both Wales and Northern Ireland in response to consultations with professionals and parents.

Visscher (2001) provides one of the most comprehensive overviews of the use of performance indicators in schools internationally. He argues that the consequences of school-level performance indicators are determined by the interaction between four broad groups of factors:

1. The *nature of the information* published, e.g. raw school performance scores versus value-added data.
2. *The way in which* the information is fed back to intended users, for example, whether it is accompanied by an explanation of what

the data mean, or whether complicated indicators are used without clear discussion.

3. *The degrees of freedom of intended users:* the nature of the local school market and whether, for example, an alternative school exists for parents if their local school does not appear to perform well.
4. *Actions of systems:* to what extent do governing systems seek to take action to correct poorly performing schools?

Visscher argues that the interaction of these four groups of factors can generate three categories of problems: technical or analytical issues around the construction and aggregation of performance indicators; usability issues related to the clarity, utility and comprehensibility of the data presented to service users; and political or societal issues, linked to the broader implications of the use of performance indicators on public service provision. We look at each of these sets of issues in turn.

Technical and analytical issues

Much of the academic study devoted to the issue of school league tables and performance indicators has focused on the numerous technical difficulties with constructing valid and reliable indicators of goals as multi-faceted as educational achievement and teaching quality. These debates have continued in the UK, as well as in the United States, Australia and Continental Europe. Goldstein and Spiegelhalter (1996) provide a comprehensive discussion of the technical issues surrounding the estimation of school effects, including value-added scores, and in particular the need to provide confidence intervals,[1] providing a variety of perspectives on both the technical and social aspects of league tables.

One of the earliest and most persistent critiques of school league tables is that, particularly in the case of raw, uncontextualised averages, league tables can end up, as Visscher (2001) puts it, revealing 'more about schools' catchment areas than about the quality of school processes

1 **Confidence interval** – When making comparisons between institutions it is assumed that we are interested not merely in how they happened to perform at the time when the data were collected, but how they compare in terms of their underlying 'effectiveness'. Thus, for example, to base a comparison using just one randomly sampled student from each school would be very unreliable and hardly acceptable. The question is then to determine how many students contributing to a school's score would be adequate. By providing a range or interval for each school we can indicate the relative accuracy for different schools, with larger intervals associated with less accuracy. Judgements can then be made about whether differences can be ascribed to chance variation due to small numbers of students, or may reflect real differences. Goldstein and Spiegelhalter (1996) provide a detailed discussion.

and performance'. The apparent solution to this problem has been the adoption of value-added scores, which attempt to adjust school mean scores by taking account, in an appropriate statistical model, of the prior educational attainment levels of the students at the time that they enter their school. The first '**value-added**' measures were introduced in England in 1998, initially contrasting the performance of pupils at Key Stage 3 with their GCSE results two years later. In 2003, the National Audit Office, on the basis of recommendations from the National Foundation for Education Research, announced that performance information should also account for 'other external influences on performance', alongside prior attainment. This led to the development of a 'contextualised' value-added measure. Launched in October 2004, the new measure particularly incorporates information on peer group measures gathered by PLASC, and is still being refined and developed. However, while generally viewed as an improvement to relying solely on raw data, the construction of value-added indicators has been shown to be subject to other statistical concerns.

Ladd and Walsh (2002), in an often quoted study, conducted a review of the effects of introducing value-added league table measures on schools in North and South Carolina. They found that, even upon the introduction of a value-added component, schools serving higher-performing students were more likely to be deemed effective than schools serving lower-performing students. The data that they used, however, is for just a one-year period between final and initial (adjustment) measures and does not generally carry over to longer periods of schooling. They also fail to study model misspecification in terms of differential school effectiveness (see below). Their attempt to study the effect of measurement error consisted of adding a prior achievement score (two years before the outcome measure) as an *instrumental variable*.[2] It is not clear what inference can be drawn from this, however, since the precision of the estimated scores will change and this will also result in changes in rank position. In fact, their result is not replicated elsewhere. Thus Ferrao and Goldstein (2009), using data from Portugal, find a very high correlation for value-added estimates with and without taking account of measurement errors, where they use external estimates of measurement error

2 **Instrumental variables and measurement errors** – Measurement errors in student test scores can distort relationships and conclusions about rank orderings, unless they are adjusted for. This is typically difficult and the use of instrumental variables is one possible approach. In this, a further measure is chosen on the grounds that it is a strong predictor of the test scores but is itself uncorrelated with any of the measurement errors. For a full discussion of this and other approaches see Ecob, R. and Goldstein, H. (1983). "Instrumental variable methods for the estimation of test score reliability." Journal of Educational Statistics 8: 223–41.

and also carry out sensitivity analyses. In addition, it has been shown (see e.g. Goldstein *et al.*, 2007) that the common value-added model, a so-called 'variance components' model, is indeed a misspecification and that the apparent positive correlation of value-added scores with raw scores can be explained as a result of this misspecification. The misspecification arises because it is generally the case that schools are 'differentially effective', that is, their rankings differ for different kinds of students such as those with high initial achievements as opposed to those with low initial achievements.

> "League tables not only fail to provide reliable information on the quality of schools as they currently operate, but they can offer even less information about future performance – the key issue for parents"

Another major issue, which is of particular relevance to school-based league tables, is that of **uncertainty**. As Visscher (2001) points out, 'even if student achievement scores have been adjusted for relevant student background characteristics … precise school performance remains uncertain as a result of large confidence intervals' (202). Large confidence intervals are just one of the results of the relatively small sample size constituted by the average school's yearly cohort. In research on this problem in the United States, Kane and Staiger (2002) find that the median elementary school has only 69 students per grade (in the UK, the average primary school year group is just 40). They point out that 'the 95% confidence interval for the average fourth-grade reading or math score in a school with 69 students per grade level would extend from roughly the 25th to the 75th percentile among schools of that size' (95). While the school results provided by the government for England, at least in the case of value-added measures, do contain such 'interval estimates', these are almost universally ignored in media presentations, and raising awareness of this issue in the media is particularly important (see Goldstein and Spiegelhalter (1996)).

Finally, as Goldstein and Leckie (2008) and Leckie and Goldstein (2009) point out, 'there is additional uncertainty arising from the fact that secondary school "league tables" are always out of date, since they refer to the performance of a cohort who began secondary schooling several years earlier … Over this period, currently seven years, the performance

of many schools changes considerably, limiting the extent to which current school performance can be used as a guide to future performance. Crucially, the league tables make no statistical adjustment for, nor do they warn about, the uncertainty that arises from predicting into the future'. Therefore, league tables not only fail to provide reliable information on the quality of schools as they currently operate, but they can offer even less information about future performance – the key issue for parents. In fact, these authors point out that when all of this uncertainty is taken into account, very few (less than 5%) schools can be differentiated in terms of their predicted value-added scores, thus making the tables effectively useless for purposes of school choice. These technical issues are important. They demonstrate the inherent limitations of comparisons based upon rankings, and for this reason they should be highlighted whenever league tables are published.

It has been suggested that one solution to the uncertainty of rankings is to publish only grouped data so that, for example, an institution can only be identified in say the bottom, middle or top group. The problem with this lies in distinguishing institutions that lie either side of a group boundary, and for this reason it seems preferable to present the full uncertainty of information in such a way that users can absorb it readily. Leckie and Goldstein (2011) propose one way of doing this based upon simulations designed to make bespoke comparisons between any chosen set of institutions. The result of this is to allow a statement about the probability that any one school is really performing better than one or more comparators. This brings us to a second category of issues: those surrounding the usability of league tables.

Usability issues

As Kane and Staiger (2002) point out, no performance measure is likely to be perfect; 'even noisy (unreliable) performance measures may provide useful information that can be incorporated into a carefully designed incentive contract. The problem resides not with the measures themselves, but with the way that these measures are often used'. The main criticisms around school league tables' 'usability' tend to regard the accessibility and intelligibility of indicators.

Visscher (2001) raises questions around **equality of access amongst parents**. He claims that 'in the United Kingdom … even after almost 10 years of publishing school league tables, a considerable percentage

of (especially low socioeconomic status) parents remains unaware of their existence' (204). He also argues that, for certain groups of parents, the structure of league tables and performance indicators is not easily interpretable. This concern is echoed by Wilson, Croxson and Atkinson (2006). In their interviews with teaching staff across the primary and secondary sector, they found that more than half of respondents believed that, while value-added measures had been a beneficial development, most parents would not engage with the new performance measures. They believed this would be due to the new complexity introduced by additional indicators, as well as the dominance of the '5 A*–C' indicator in the public psyche (2006, 166).

> "The problem resides not with the performance measures themselves, but with the way that these measures are often used"

Goldstein and Leckie (2008) also highlight the problems associated with the multi-purpose nature of school league tables. They argue that, despite any possible value for accountability purposes, the use of *contextual* value-added rankings is of little use for parental choice: 'the relevant question for a parent is whether, given the characteristics of their child, any particular school can be expected to produce better subsequent achievements than any other chosen school or schools. If a school level factor is associated with achievement this is strictly part of the effect being measured and therefore not something to be adjusted for' (68). This, coupled with the uncertainties of gauging future school performance from current data, severely diminishes the utility of league tables for informing parental choice. This is also stressed in the Leckie and Goldstein (2011) proposal for comparisons.

The second group of users, on whom less emphasis is generally placed, is that of schools and teachers themselves. There is currently a great deal of scepticism amongst teaching professionals as regards the expanding role of league tables and performance monitoring (Wiggins and Tymms, 2002). Teachers working in areas of high social and economic disadvantage in particular often feel that, even with more contextualised data, league tables do not often provide an accurate reflection of institutional quality. Furthermore – given the centralised nature of the English national curriculum – schools may have only limited possibilities to re-organise themselves more effectively.

Applying the lessons: the example of Hampshire

In an experiment in one English local authority (Hampshire) in the late 1990s, value-added estimates were introduced for primary schools that were not published, but utilised by the authority and head teachers as a 'school improvement' tool. To this end, the detailed yearly scores were fed back to schools as one item of information within an inspectoral system so that it could be used alongside other information. This use of value-added estimates as a 'screening device' has the potential to avoid many of the harmful side effects of published tables while still retaining key elements of an accountability device. We know of no other similar attempt within education to move in such a direction (Yang *et al.*, 1999).

Political, ethical and societal issues

Finally, and perhaps most significantly, a number of studies have sought to assess the broader implications of school-based league tables for educational objectives and outcomes. Experiments have revealed that the introduction of performance monitoring can quickly shift managerial focus to consider their league table position, even at the expense of overall performance (Keasey *et al.*, 2000). An extensive critique of school league tables using both data sources and case histories is given by Mansell (2007). The literature on the development of a league table culture in schools has highlighted a number of adverse consequences, including attempts to 'game' the system, adverse effects on staff morale, and, in some cases, the incentivisation of behaviour which may actually prove detrimental to educational outcomes.

One outcome of high-stakes, publicised indicators is their potentially adverse effect on teaching staff. In their study on the consequences of school league tables, Wiggins and Tymms (2002) carried out interviews with teachers which revealed growing pressure on staff and the frequently detrimental effects on morale. Comparing a league table-focused educational culture in England with that in Scotland, they find that the stress of performance targets is increasingly associated with a more 'short-termist' approach among English teaching staff and, in some cases, the development of a blame culture (46). They conclude that 'high-stakes, single-proxy indicators, particularly when presented in league tables, can have significant dysfunctional effects' (47). Visscher (2001) also highlights the institutional damage done by 'naming and

shaming' schools persistently at the bottom of the league tables. He argues that presenting league tables as a simple comparative measure will always lead to some schools performing at a relatively lower standard, but that the focus should remain on whether each school reaches those standards considered appropriate.

While the case can be made that the pressure generated by publicised league tables forces schools to drive up standards, one of the less expected findings of Wiggins and Tymms' survey was that teachers in Scottish primary schools (whose results are not publicised in league tables) felt under greater pressure to meet performance targets than teachers in England (45). What's more, schools deemed by performance monitoring to be 'good' were just as likely to find performance indicators problematic as 'poor' schools, and there was agreement across both nations that external, standardised performance indicators were not particularly good at judging overall performance and that internal systems controlled by schools themselves would be more effective.

In the United States, the use of league tables as a measure of teachers' performance is even more pronounced. Ranking is increasingly being used to judge individual teachers and results may even link to salary and promotion prospects. Thus, for example, in August 2010 the *Los Angeles Times* published a league table for 6,000 teachers with few attached caveats, based upon single-year measures of progress. The No Child Left Behind legislation from 2001 mandates states to reach strict targets every year using standardised tests and this has encouraged the proliferation of league tables. Value-added tables are used increasingly, although few attempts have been made to provide confidence intervals, even though these will be substantially wider than those for schools due to the smaller numbers of students involved. Newton *et al.* (2010) emphasise the difficulty of drawing causal inferences about individual teachers when students encounter multiple teachers across time, and they point out that there are large amounts of uncertainty, reflected in instability over time.

As external pressure on teachers to meet performance targets and maximise league table rankings increases, many authors also discuss the growth of techniques linked to 'gaming' the system (Wilson, Croxson and Atkinson, 2006; Wiggins and Tymms, 2002; Visscher, 2001). In some cases, these studies argue, institutions become so focused on the measures and standards employed by league tables that they begin to deliberately manipulate their data or behaviour to produce the desired results, regardless of potentially adverse effects.

Smith (1995) sets out a number of means by which 'gaming' takes place:

- concentrating on those students with whom most 'profit' can be gained to improve a school's Student Progression Information (SPI) while ignoring the needs of students at either end of the ability spectrum;
  - selective student admissions;
  - removing 'difficult' students;
- concentrating on examination performance to the exclusion of other qualifications;
- the confusion of correlation with causation in interpreting school performance data;
- 'creative reporting' of data;
- teaching for the test; and
- depression of baseline/intake test scores to improve the value-added scores.

Such techniques are common responses to the use of high-stakes indicators, but may well subvert the original intentions of performance monitoring, crowding out genuine efforts to raise standards. Most importantly, there is evidence to suggest that the results of such practices may in some cases actually prove detrimental to overall educational standards.

Wilson, Croxson and Atkinson (2006) carried out interviews with a variety of teachers and headteachers. Many schools reported that they did tend to focus extra resources on 'borderline' pupils (those who are likely to achieve C or D grades). This was acknowledged to have consequences for others; one interviewee admitted 'the bright kids still prosper … I don't think they miss out at all. But I think the lower ability ones potentially do' (164). Others reported that they deliberately shifted these borderline pupils to vocational qualifications; according to one headteacher, '… we started last year, we introduced a GNVQ course which is a double award and we deliberately targeted that at the middle of the road pupils, those pupils who might get four A to C passes' (163). This practice, while potentially boosting a school's results, has been shown to have negative consequences for the pupils themselves. Research by Robinson (2001) on qualifications and wage premiums found that, while there is no premium linked to holding lower level NVQs, possessing lower grade GCSEs is associated with a modest wage premium: on average, 9% for men and 5% for women.

Evidence from the United States also raises questions about the educational benefits of high-stakes performance monitoring in schools. In the state of Texas, under former Governor George Bush, a very high-profile testing programme was instituted in 1990 for grades 3–10 (ages 8–16) in Texas schools. The results are used to rank schools in league tables and certain funds are allocated on the basis of the test results. Over the 1990s very large gains in student test scores were observed, and certain ethnic minority differences were reduced. Dubbed the 'Texas miracle', these results have been used as a justification for such testing programmes involving rewards given to schools for performance on the tests. The most important manifestation of this trend in the US is the 'No Child Left Behind' Federal Education Act of 2001 (www.nochildleftbehind.gov/) which mandates testing of all school pupils in grades 3–8 (ages 8–14) and publication of results in league table form. In one important respect it goes further than legislation in England by giving parents the right to transfer a child from a low-scoring school to a higher-scoring one.

However, researchers from the RAND corporation have compared the results of the intensive testing programme in Texas with results obtained from a national testing programme, the National Assessment of Educational Progress (NAEP) that is carried out over the whole of the US (http://epaa.asu.edu/epaa/v8n49/). What they found was that for mathematics and reading, compared to the rest of the US the comparative gain in test scores over time of the Texas students on the national test was much less than that implied by the Texas test scores, and in some cases no different at all from changes found in the US as a whole. Moreover, the ethnic results from NAEP showed that, if anything, in Texas the differences were increasing rather than decreasing. The researchers conclude that the concentration on preparation for the Texas state tests may be hindering an all-round development of mathematics and reading skills, especially for minority students.

The appropriate **role and response of governments** in this area has been much disputed. Governments' desire to foster greater accountability within public services, as well as to allow a wider scope for user choice, has been central to the growth of league tables and performance indicators for schools. However, a number of studies have been critical of governments' lack of responsiveness to the challenges posed by league tables. Kane and Staiger (2002) highlight psychological findings suggesting that people tend to be overconfident in predicting future performance on the basis of current performance (Kahneman and Tversky, 1971). This may lead governments to 'draw unwarranted conclusions on the effectiveness or ineffectiveness of policies based

upon such short-term fluctuations in performance' (102). This is reinforced by the findings of Leckie and Goldstein (2009), who show that past performance is poorly correlated with future performance.

A number of countries, including Denmark, Scotland and Northern Ireland, have used these arguments in deciding not to publish school performance indicators. After a public consultation on the future of school league tables, the former Northern Ireland Education Minister, Martin McGuinness, chose to switch to a system in which secondary schools provided their own information on exam results to parents. Speaking after the decision, McGuinness stated,

> *'Many respondents felt the tables were divisive and failed to offer schools the opportunity to give parents a rounded picture of the school. Overall there was a majority in favour of replacing the tables. I am convinced that this decision is the right one for our schools and our parents. In future schools will be able to set their performance in the context of the school as a whole'*
> (Russell, 2001)

Many governments continue to believe in the value of performance monitoring in the education sector and some have recently decided to embrace league tables. A notable example is Australia, which, in 2010, introduced a website (www.myschool.com.au/) which aims to provide parental choice using simple, non-value-added rankings of test scores.

The challenge for future policy in this area will be focused on successfully identifying and finding means to address some of these problems – an area that forms the focus for this report.

Issues for further research

More research is needed on the actual effects of league tables on the behaviour of parents, schools, policymakers and inspection systems such as Ofsted. Some of this already exists, but much is anecdotal. In particular, it is important to gain more evidence on how institutions 'game' the system, the extent to which this distorts other educational aims and ways in which this can be mitigated.

Summary of issues and recommendations

The studies cited here have highlighted a variety of technical, political and social issues linked to the growth of league tables and performance monitoring in schools.

A summary of these issues and recommendations to address them are below:

- The 'measurement response' principle, sometimes referred to as 'Goodhart's Law', typically operates to change the behaviour of an institution when it knows that it is being monitored and that its future may depend on the results of that monitoring. Although it is often argued that this is precisely the point of league tables, the evidence is that they could have unintended negative effects.
- The linking of league tables to rewards should be weakened to reduce the side effects of inappropriate 'gaming' and to reduce stress among teachers, parents and students. This would also have the desirable effect of making the results a more objective evaluation of performance. The problematic consequences for schools serving the most disadvantaged pupils particularly need to be addressed.
- The government should consider ways to prevent league tables being exploited by the media, such as encouraging measures of uncertainty to be provided around any institutional results. Associated with this there could be a campaign to better inform the public at large about the strengths and limitations of league tables, although any such attempt poses considerable challenges.
- Consideration should be given to alternative ways of using quantitative information to monitor educational performance generally. This can be achieved by in-depth study of a sample of schools and students within a national database. A useful model is the Assessment of Performance Unit that was set up in the 1970s in England and discontinued in the 1980s (Gipps and Goldstein, 1983).
- Consideration should be given to using performance information as a screening device rather than publishing as league tables, as in the Hampshire experiment. This could be accompanied by an emphasis on evaluation and inspection systems that are designed to emphasise ways of assisting schools to cope with problems rather than 'exposing' them using public rankings.
- Ways to rely less on a small number of indicators should be sought, as well as those which cover more aspects of learning.

- More appropriate statistical analysis models should be used to describe institutional differences that allow for differential performance for different groups of students. In particular a shift away from the comparison of individual institutions towards research that helps to identify modifiable factors that appear to be related to good performance.
- It is important to guard against harm from the short-comings of school performance measures. An ethical code to govern their publication should be formulated, as suggested by Goldstein and Myers (1996). This would be based on two broad principles: that unjustified harm to those to whom the information applies should be prevented, and that there should be no *absolute* publication rights for performance data.
- Further consideration needs to be given to the role of inspection and accreditation agencies as a means of evaluating individual institutions. Trust in such agencies may not be easy to achieve, especially when they are perceived to be instruments of government. A discussion of such agencies is given in Appendix A.

2 Higher education

Alongside the ongoing debates around the growing 'league table culture' in schools, this issue is becoming increasingly salient in the field of higher education. A number of developments in recent years have fuelled the rise of higher education performance monitoring:

- The university sector in the UK has expanded rapidly to include former polytechnics, creating a much more diverse higher education sector.
- Tuition fees have been introduced, and have risen dramatically over the last 10 years; this shift of the funding burden from state to private individual has given rise to the idea of the 'student consumer'.
- Finally, the higher education sector is increasingly a globalised one. Universities across the world are now in competition for students, staff and funding; between 1999 and 2009, the number of students attending a university outside their home countries rose 57% to three million, and half of the world's top physicists no longer work in their home countries (Baty, 2010). The Higher Education Funding Council for England (HEFCE) (2008) reports that international students, as well as foreign governments and scholarship bodies, are increasingly using league tables to inform their decisions.

All of these developments have combined to create an increasingly marketised higher education sector, in which prospective applicants are confronted with a great deal of choice and invest a significant amount of money. Education markets, like all markets, need information to function effectively. The growing demand for information on higher education institutions means that university league tables have become a ubiquitous feature of academic life.

Higher education league tables also confront a number of challenges which are less evident in the school sector. The first is the relative independence of higher education institutions. The government's role in defining the nature and purpose of league tables in this sector is far less

evident; this is a task generally taken on by the media, which can sometimes create problematic consequences. The second is the diversity within the sector, both through the lack of standardised qualifications and the variety of different 'missions' which academic institutions now define for themselves. In addition, particularly for international comparisons, there is no common baseline measure of prior achievement that would allow value-added comparisons.

Vaughn's (2002) review of commercial league tables for universities in the United States defines three broad types of ranking system:

1. The first category involves aggregating a variety of different indicators to create rankings based on an 'overall' single score for each institution. Notable examples include *The Times* and *The Sunday Times* university rankings in the UK.
2. The second type is divided according to subject area, so that the comparison is between different departments, rather than different institutions.
3. The third category Vaughn simply defines as 'other', pointing out that the proliferation of ranking systems over the last few years means that a variety of new types have emerged. Vaughn gives the example of the Recruit Ltd. approach in Japan, which ranks institutions simply by each of the 88 questions in its survey.

Given the growing importance of university rankings, as well as the diversity of their compilers, aims and ranking mechanisms, the impact of league tables on higher education has become increasingly controversial.

Technical and analytical issues

There are two recurrent problems raised with regard to the technical aspects of constructing higher education league tables. The first relates to the **types of indicators used** and the second to the **scaling and weighting** of these indicators. On the types of indicators employed, measures used are frequently identified as badly chosen, poor proxies, or lacking internal construct validity.[3] In their comprehensive review

3 **Construct validity** – Essentially this is an assessment of how well a test really measures what it claims to measure. There is a very large literature on this topic and a recent comprehensive discussion can be found in Lissitz, R. W., Ed. (2009). The concept of validity. Charlotte, NC, Information age publishing, INC.

of higher education league tables, HEFCE (2008) argued that many measures were determined largely by the data currently available, rather than a clear or coherent concept of academic quality. Oswald (2002) has questioned a common focus on institutional spending and resources, arguing that it might be unclear from an economist's perspective why such high spending should be uncritically rewarded, rather than investigated on a value-for-money basis. Additionally, Bowden (2000) highlights the fact that a number of the variables used are under the control of the institution itself (for example, the proportion of first class degrees awarded). This can create adverse incentives. HEFCE point out that 'data require interpretation and some conceptual framework, but league tables often combine performance indicators in an ad hoc way that may not even reflect the compilers' own concept of quality' (12).

> "Part of the problem of university rankings is that it is impossible to reach a universally acceptable definition of the concept of academic 'quality'"

Part of the problem stems from the fact that it is impossible to reach a universally acceptable definition of the concept of academic 'quality'. University rankings have been variously criticised as either overly dependent on one particular aspect of academic quality, or as unable to demonstrate how the variety of indicators they employ serve to illustrate the concept. For example, Yorke (1998) analysed *The Times'* league table and found that 93% of the variance in institutional scores could be explained by the research variable alone. A number of other commentators highlight the lack of internal construct validity. Berry (1999) examined the discrepancies between different higher education rankings by comparing rankings in *The Times* and the *Financial Times*. He argued that, if the tables created a reliable picture of the quality of British higher education, discrepancies between them would be few and minimal. His results, however, showed large differences in the rankings of different institutions; almost 30% of the institutions were ranked 10 places apart or more, and 14% at least 15 rank places apart in the two tables, while one was placed 20 ranks lower by the *Financial Times* than by *The Times*, and four others over 20 places higher (7). Clarke (2002) raises the question of the correlations between indicators, posing the question of whether high correlations are a sign of validity because all indicators are measuring academic quality, or redundancy (all indicators measuring the same type of academic quality).

However, over the last few years it does appear that there is greater consensus between rankings on the types of indicators considered important. Dill and Soo (2005) compared five different higher education rankings; two used in the UK (*The Times* and *The Guardian*), and three others used in the US, Canada and Australia. Their analysis of these rankings suggests that 'a common approach to measuring quality in higher education is emerging internationally … We can observe that input measures have a prominent role in all five rankings and that the input measures used are quite homogenous' (499). Output measures, on the other hand, are subject to greater discrepancies; some rankings focus on the graduation rate, while others look at graduate employment opportunities or graduate satisfaction with the programme. They also find that input measures are often weighted at the expense of the teaching and learning process. This creates problems for prospective students using league tables; Dill and Soo believe that rankings tend to 'reflect the universities' recruitment policies instead of the actual quality of education' (510). Furthermore, they argue that longitudinal surveys have revealed that the learning environment and student involvement in the learning process are the variables which have the greatest impact on student outcomes. Yet these types of process measures are generally missing in reviewed league tables. Terenzini and Pascarella (1994) also found that the quality of teaching and the quality of research appear to be largely independent of one another.

As well as the types of indicators used, the **scaling and weighting** of different indicators is a further area of controversy. Several commentators find the weightings of indicators to be 'questionable' or 'lacking convincing rationale' (Bowden, 2000; HEFCE, 2008). Dill and Soo (2005), in their review of a number of different ranking mechanisms, found that the five reviewed tables 'fail to provide a theoretical or empirical justification for the measures selected and the weights utilised' (506). For example, they highlight the graduate unemployment measure commonly used in UK league tables; analysis of the data revealed no statistically significant differences between most UK universities in patterns of graduate employment. They concluded that a more useful indicator might be data on whether students are employed in graduate-level jobs. Evaluating different weighting systems via sensitivity analyses is a useful procedure. If rankings change markedly when weights are changed then this would seem to pose a serious problem for interpretation.

Finally, HEFCE's (2008) report highlights the fact that higher education institutions do not feel they have sufficient influence on the compilers and the methodologies used in the rankings. For example, many

would favour more emphasis on a value-added approach. In the case of higher education institutions there may well be comparable inputs domestically, for example A level grades in England, but this looks less feasible for international tables. Additionally, as Dill and Soo (2005) point out, the types of process and output measures which may be of most interest to prospective students, such as teaching quality and graduate job prospects, are also often neglected. However, most institutions acknowledge that the growth of league tables has prompted them to implement better data collection systems.

More recently, the weekly magazine *Times Higher Education* has revised its previous ranking policy and begun to produce a variety of international rankings based upon surveys of students and staff as well as research publications, drop-out rates and academic results. As well as providing information for prospective students, one of the stated aims is to provide information to university administrators.

Usability issues

Following the diversification of the student body through the rapid expansion of higher education in recent decades, questions about the usability of league tables have gained increased attention. Compilers of league tables face a challenge: they must attempt to strike a balance between encapsulating as much relevant information as possible, and making the data comprehensible. At present, the solution most commonly opted for is the collection of data according to a number of different measures, and then the aggregation of all of these indicators into an overall score. This has the advantage of creating a single, easy to grasp ranking. However, it is not a method without its drawbacks, and the limitations of this approach are becoming increasingly apparent in an era of growing student diversity.

Turner (2005) refers to this method of aggregation as 'excessively simplistic', arguing that simply adding indicators together does not allow for a robust or meaningful ranking system. HEFCE (2008) also points to large discrepancies in the indicator weightings used: 'entry standards' are given an 11% weighting by *The Times* but a 23% weighting by *The Sunday Times;* the 'staff-student ratio' is weighted at 17% by *The Guardian* but only 9% by *The Sunday Times* (16). Although they can have dramatic effects on an institution's ranking, these discrepancies are often not made clear; HEFCE (2008) claims that there is 'insufficient transparency' about

the way league tables are put together and that compilers need to make clearer distinctions between inputs, processes and outcomes.[4]

In light of this increasing diversity, a number of commentators question whether the indicators being used are really the most informative for prospective students. One potentially positive step is the development of the Student Satisfaction Survey in the UK, which questions final year undergraduates about their views on their course (www.thestudent-survey.com/). It is not yet clear what benefits this survey may have, but such efforts may help to diversify indicators in a way which better caters to the concerns of prospective students.

Bowden (2000) expands this point, arguing that diversity of students should be considered alongside diversity of indicators. Increasing numbers of 'non-traditional' students may hold a very different set of priorities. HEFCE (2008) claims that the five university league tables they reviewed 'do not provide a complete picture of the sector, with a focus on full-time, undergraduate provision and institutional, rather than subject-based, rankings. This excludes a wide range of specialist, postgraduate, small or predominantly part-time institutions' (5). A number of authors agree, arguing that league tables are ill-equipped to cater for the interests of postgraduate students, those pursuing interdisciplinary studies, mature students, part-time students, local students, overseas students, those entering with alternative academic qualifications and those studying for qualifications other than a degree (Eccles, 2002; Sarrico *et al.*, 1997 in Bowden, 2000; HEFCE, 2008). This has led to the neglect of a number of indicators which may be of interest to these new, non-traditional types of student, such as the percentage of part-time students, the flexibility of choice in programme construction, the availability of distance learning programmes, living costs, bursaries and non-academic facilities (Bowden, 2000; HEFCE, 2008).

HEFCE (2008) found that many universities believed that 'traditional' prospective students (younger applicants with higher academic achievement and social class) are now more likely to use league tables than their less traditional peers. A UK survey on student choices (Connor *et al.*, 1999 in Dill and Soo, 514) found that the most important fac-

4 Appendix C of HEFCE's 2008 report shows the correlations between the overall scores given by different ranking systems. While the correlations between the British rankings systems are relatively high – such as that for *The Guardian* and *The Times* tables (0.88, T3.5.1) – the correlation between the top 100 or so institutions in the Academic Ranking of World Universities (AWRU) and Times Higher Education (THE) rankings is only moderate (0.66, T3.8.1).

tors influencing the choices of university applicants were: the course, academic quality (particularly teaching reputation), entry requirements, employment prospects, location, academic and support facilities and cost of study. Many of these are downplayed or neglected entirely by traditional league tables. What's more, it appears that this mismatch between indicators employed and those issues of most concern to prospective students may be restricting the use of league tables as a decision-making tool. In the UK, a 2007 UNITE Student Experience survey questioned 1,600 students and found that 29% mentioned university league tables as a factor influencing their decisions (HEFCE, 12).

However, it should be noted that those compiling league tables face a difficult trade-off between the complexity required to capture this more diverse range of indicators and the extent to which tables appear accessible to their target audience. The re-launch of the *Times Higher Education* (THE) tables has been one of the most recent examples of attempts to navigate this trade-off. The tables use 13 separate indicators with weight given to citations, institutional research income and surveys of 'reputation'. The results are published in five broad groupings as well as an overall index. Individual disciplines are also examined and the rankings are supplemented by critical commentaries. The *THE* claims that these tables are 'built on a sophisticated range of metrics rather than opinion'. Whether an increased 'sophistication' can be equated to a greater usefulness, however, is not entirely clear. This is especially the case since the usefulness of quantitative measures such as citations has increasingly been questioned (see for example, Volume 24, issue 1 of the journal 'Statistical Science', 2010, devoted to this topic).

Political, ethical and societal issues

The dramatic changes to the structure of the higher education sector in the UK which have taken place over recent decades have raised questions around the possible political and social repercussions of a dominant culture of rankings. As discussed, neglect of diversity is a frequently raised problem; HEFCE (2008) claims that 'there is an enduring reluctance among UK compilers to distinguish between institutions with different missions and compare like with like' (56). As such, current higher education rankings 'largely reflect reputational factors and not necessarily the quality or performance of institutions' (5). Entry qualifications, good degrees and Research Assessment Exercise (RAE) grades are more highly correlated with total scores than other indicators, such as National

Student Survey results and teaching quality scores. There is also the idea that rankings detract from the educational process, fostering an **instrumentalist approach** that reduces higher education to a uniform product.

Most problematic of all, as league tables take on increased significance and greater meaning, they may begin to promote perverse behaviour within higher education institutions. According to HEFCE (2008), they often encourage universities 'to take superficial actions to improve their positions rather than engaging in the more challenging task of enhancing teaching and student learning'. (12) There is growing evidence that 'gaming' of the league table system is now relatively widespread, particularly in countries where league tables are a deeply-rooted aspect of the higher education sector.

> "Current higher education rankings 'largely reflect reputational factors and not necessarily the quality or performance of institutions'"

Dill and Soo (2005) discuss a number of instances of this behaviour occurring in the United States. For example, at Cornell University, because the proportion of alumni who make donations to an institution is viewed by some US league tables as a measure of graduate satisfaction, Cornell was shown to have decreased their numbers of alumni by eliminating those for whom they did not have a valid address, and those who had attended but not graduated. In some cases, US institutions have now made the SAT (entry qualification) test an optional requirement for applicants, since this would ostensibly boost their average reported entry score (516). Dill and Soo add that 'what is conspicuously missing in all these reports of college and university response to US league tables are active efforts to improve teaching and learning for students' (517).

Baty (2010) provides further examples of the problem of perverse incentives – perhaps more worryingly, in less developed countries attempting to enter the higher education 'market'. He points out that 'the new weight placed on an international dimension in global rankings has led some institutions to indiscriminately and rapidly recruit international faculty and students, which is detrimental to local talent'.

Given that many governments are keen to develop and expand the higher education sector, the gaps between the priorities of league tables and those of politicians will need to be addressed. Institutions

now 'need to manage the tensions between league table performance and institutional and governmental policies and priorities (e.g. academic standards, widening participation, community engagement and the provision of socially-valued subjects)' (HEFCE, 6). Additionally, as higher education becomes more competitive, internationalised and expensive, we are likely to see the influence of league tables increase. According to HEFCE (2008), they are now 'being used for a broader range of purposes than originally intended and being bestowed with more meaning than the data alone may bear' (7).

The general consensus appears to be that league tables are likely to remain a prominent and increasingly influential aspect of the higher education landscape. Universities themselves are therefore compelled to take their rankings seriously, despite concern about the quality of the information currently on offer. The question now is to what extent higher education ranking systems can be improved to better reflect the demands of prospective students and to minimise the unwanted side effects they are sometimes capable of generating.

Identifying solutions

The last few years have witnessed a number of attempts to identify solutions to the problems discussed above. The UNESCO European Centre for Higher Education Policy (UNESCO-CEPES) has been instrumental in convening meetings and expert panels to discuss the challenges surrounding higher education rankings. In June 2002, it convened a three-day meeting in Warsaw, Poland to discuss the methodologies of ranking systems, involving 50 experts from 12 countries. In 2006, in conjunction with the Institute for Higher Education Policy in Washington, they also formed a working group which developed the Berlin Principles – a set of guidelines on quality and good practice for higher education rankings. Some of these recommendations are discussed below.

Policy issues

There have been a number of suggestions for improving the political and social impact of league tables. This is widely considered to be the most challenging aspect of public sector performance management. The Warsaw meeting stressed the difficulties – as with many publicly-provided services – of identifying the true purposes of higher education. They

felt there was value in independent analyses of what higher education constitutes and what its aims are. They also felt that ranking systems should be just one of a more diverse range of approaches to assessing performance in the higher education sector; this need for alternative approaches is also stressed by Vaughn (2002).

According to the Berlin Principles, rankings should both recognise the diversity within institutional 'missions' and seek to complement the goals of governments and other bodies overseeing higher education. Dill and Soo (2005) also point to the important role of league tables in encouraging universities to make changes that actually improve the quality of teaching and student learning. International rankings also need to take account of the features which are specific to particular higher education systems, and recognise that there are various distinct notions of quality.

It is also important to try to contextualise measures. In the case of student performance this could utilise intake achievement at least within systems, similar to the value-added analysis used for schools, although the lack of comparable outcome scores when using degree classifications renders this problematic. Clarke (2002) suggests that statistical uncertainty should be conveyed in an attempt to address the problem and this is an area where more work is needed to ascertain how this can be done to best effect.

Finally, while taking its core users to be prospective students, the CHE also aims to account for the informational needs of institutions themselves. They therefore offer detailed analysis of the student survey for single departments, with data that goes beyond the published indicators. Federkeil suggests that this has been an effective approach, pointing out that the CHE now receives much positive feedback, even from departments that come off badly. He argues that the detailed feedback provided helps poorly-performing institutions to analyse problems and initiate reforms.

Applying the lessons: the Carnegie Classification

The Carnegie Commission on Higher Education was established in 1967 by The Carnegie Foundation for the Advancement of Teaching. It was set up to study and make recommendations on the major issues facing US higher education, but became most famous for its attempt to develop a new classification scheme for universities, designed to meet the analytic needs of those engaged in research on higher education.

A key goal of the new system was to call attention to the considerable institutional diversity within US higher education. The classification grouped roughly comparable institutions into categories designed to enable researchers to make meaningful comparisons. As McCormick and Zhao (2005) put it, institutions were grouped 'according to what they did and who taught whom. Operationally, this was achieved by looking at empirical data on the type and number of degrees awarded, federal research funding, curricular specialisation, and (for undergraduate colleges only) admissions selectivity and the preparation of future PhD recipients. The result was a classification organised by degree level and specialisation: doctorate-granting universities, master's-level institutions, undergraduate liberal arts colleges, two-year colleges, and specialised institutions' (52).

This was a system which was widely adopted by the higher education research community, and, according to McCormick and Zhao, 'soon became the dominant – arguably the default – way that researchers characterized and controlled for differences in institutional mission' (52). However, such developments in the research community have not readily spread into the realm of commercial rankings.

Applying the lessons: the example of Germany

Another means of identifying possible recommendations is to look at the approaches other countries are taking in addressing these issues. Higher education league tables in Continental Europe are a much more recent, and therefore less embedded, addition to the higher education landscape, particularly in comparison to the UK and the US. One consequence of this situation is that European countries may be able to learn lessons from the common drawbacks of the US/UK approaches and come up with systems which circumvent them. Federkeil (2002) highlights the development of league tables in Germany as one example in which many of the recommendations discussed above appear to have been taken on board.

University league tables in Germany, rather than compiled in the media, are dominated by the Centre for Higher Education (CHE) – an independent think tank. The CHE (2010) defines itself as a 'reform think tank for higher education', which aims to develop models for the modernisation of higher education systems and institutions in dialogue with decision-makers from higher education and politics. The use of a think tank as the primary compiler of league tables may avoid some of the more sensationalist reporting of rankings which takes place when they are primarily put together by media outlets, as in the US and the UK.

According to Federkeil, the CHE rankings are specifically oriented towards school leavers and therefore aim to incorporate many of the features that offer them the clearest and most useful information. Numerical rankings thus only apply to specific subjects. Overall rankings allow users to decide on the weightings for each indicator and those published by CHE simply arrange all universities in three groups, with those scoring highest in the top group, those scoring lowest in the bottom and all others considered intermediate. The grouping procedure varies according to two kinds of indicators. Facts (for instance, staff–student ratio, number of publications) are grouped according to quartiles, with the highest and lowest ranking as top and bottom respectively. In the case of subjective indicators based on survey data, the procedure takes into account the diversity of judgments within universities compared to the overall score. A university is ranked top if the confidence interval of the mean is completely above the overall mean of all universities. At the other extreme, a university is ranked bottom if its confidence interval is completely below the overall mean.

Federkeil also points out that these efforts to improve accuracy and usability have boosted the popularity of the rankings amongst prospective students; he finds that, despite the fact that rankings are a relatively recent phenomenon in Germany, one third of German students now use the tables. They also appear to have become relatively influential; Federkeil highlights the subject of psychology, which was first included in the CHE ranking in 2001. A year later, the number of applications for admission to the recommended universities increased notably, even while they remained stable overall.

Issues for further research

As regards areas for further research, there appear to be some gaps in our understanding of the way in which prospective students engage with league tables. While there exists some survey evidence on the proportion of students consulting rankings, as HEFCE (2008) points out, there is little information on the *way* in which prospective students use the tables, the *extent* to which they influence their decisions and the *types* of students who currently use them.

The second area of interest is likely to be the growing international dimension to league tables in higher education – both in terms of the way in which foreign governments and scholarship bodies interpret domestic rankings and the way in which internationally-comparative league tables are compiled. Federkeil (2002) argues that the appropriateness of indica-

tors remains heavily dependent on national higher education systems and that many of the indicators with important bearing on higher education in one country may make almost no sense in another.

The availability of new sources of information on higher education, particularly via the internet, will have an important effect on future rankings. The rise of comparative sites such as Unistats, the new flexibility in presentation and indicator weightings allowed for by online rankings, and the use of social networking sites in decision-making are also relatively unexplored areas.

Finally, research into potentially perverse side effects is required. While there is evidence in the school sector about these, there is little available for higher education. Of particular concern is the effect on socially disadvantaged students and those from ethnic minorities. It is important to know if and how universities may change their admissions policies or their awarding systems to optimise their rank position.

Summary of issues and recommendations

We have highlighted a variety of technical, political and social issues linked to the growth of league tables in higher education. A summary of these issues and recommendations to address them are below:

- Indicators need to be selected according to validity, rather than availability as currently tends to be the case. This implies more qualitative and process indicators, although care needs to be exercised in terms of their subjectivity.
- Disaggregated indicators are important and the temptation to aggregate into one index, or even a small number of indexes should generally be discouraged.
- As with schools, the uncertainty needs to be displayed.
- For users, broad categories rather than precise rankings are to be preferred and sensitivity analyses with different weightings to components should be conducted to test stability.
- Subject-based rankings should be emphasised.
- As with schools, but to a greater extent, further consideration needs to be given to the role of inspection and accreditation agencies as a means of evaluating individual institutions. Trust in such agencies may not be easy to achieve, especially when they are perceived to be instruments of government. A discussion of such agencies is given in Appendix A.

3 Crime and policing

The final area encompassed by this review is that of the impact of performance measurement on policing. As a public service, policing shares a number of significant features with education; these are areas in which aims and processes remain very difficult to define. For both services, responsibility for provision is shared between a number of stakeholders, with inspection bodies (Ofsted and HMIC), government departments (the Department of Education and the Home Office) and local authorities all invested in decision-making. Both have come under an increasingly stringent system of performance measurement in recent decades, but in both cases the effectiveness of this approach has remained the subject of contention.

> "League tables and performance indicators for the police have not gained the same public prominence as those for schools"

However, there are also important differences between these two areas, which affect the nature of performance measurement. The issue of user choice has been a vital driver of the growing use of league tables within the education sector. For police forces this has not been as significant an issue; while advocates of performance monitoring have been keen to foster an increased sense of public accountability within the police force, members of the public do not possess the ability to 'choose' a police force, although the issue of electoral accountability within the police has been raised by the new (2010) UK Coalition Government, as will be discussed. This has meant that league tables and performance indicators for the police have not gained the same public prominence as those for schools.

The development of performance monitoring for both services has taken a similar trajectory. As in education, the perceived need to find

an accurate means of measuring and monitoring police performance arose under the Conservative administration in the early 1990s. As Collier (2006) sets out, the process was triggered by a 1993 White Paper on police reform, which proposed that government should be setting key objectives regarding crime and measuring police performance against them. These objectives were initially set out in 1994, but, from the beginning of the Labour administration in 1997, have been subject to frequent alteration. According to Ashby (2005), the Labour government developed 'an increasingly tough performance management culture' (418). However, many of their initiatives also attempted to address the numerous perceived problems with performance measurement. Following the 2000 Spottiswoode Report, a Policing Performance Assessment Framework (PPAF) was developed, which aimed to create more direct and transparent measurements of performance, balance national and local priorities, and reduce the number of indicators used. The PPAF was based on six 'domains': citizen focus, reducing crime, investigating crime, promoting public safety, providing assistance and resource usage. Within these areas, 14 indicators were in operation, broadly linked to: inputs (public demand for police), processes (police behaviour) and outcomes (arrests, detections and public satisfaction).

The Police Performance Monitoring Framework continued to be developed in the later years of the Labour administration. The Police Reform Act of 2002 required the Home Secretary to produce an annual National Policing Plan, setting out strategic priorities (this has since been dropped by the new Coalition Government). According to Collier (2006), after 2004 there was also a shift away from process indicators in favour of a focus on police outputs – often perceived as easier to quantify. Fears around the centralised nature of Labour's early reforms also led to attempts to revive neighbourhood policing strategies and to engage better with local communities.

By the end of the Labour administration, police performance monitoring was dominated by *four main elements*: Public Service Agreements, National Policing Plans, the Police Performance Assessment Framework and Statutory Performance Indicators. The main *indicators of performance* were: measures of response time to emergency calls, levels of crime and rates of detection, traffic incidents, complaints against the police, cost per head of population and surveys of public satisfaction. While police forces do receive an overall performance rating, the Home Office has consistently rejected the use of league tables to summarise performance indicators, owing to the incomparability of police perfor-

mance in regions with very different social-demographic compositions and urban-rural structures.

It is this dimension of accountability to the public which looks likely to drive the next phase in the development of performance measurement in the police service. The Home Office recently set out the priorities of the new administration in their consultation document, 'Policing in the 21st Century' (2010). The chief criticism of previous performance monitoring frameworks is that 'targets and standards in policing were driven by Whitehall rather than the public'. To foster a culture of public accountability in the police force, the paper argues for the election of local Police Commissioners and much more extensive availability of data. The commitment is that, from January 2011, 'we will ensure that crime data is published at a level which allows the public to see what is happening on their streets and neighbourhoods. We will require police forces to release this data in an open and standardised format that would enable third parties to create crime maps and other applications that help communities to engage and interact with their local police force'. The aim of more transparent performance information is clearly centred on accountability; 'the increased provision of accurate and timely locally focused information to the public will be critical in empowering them to effect real change in their communities …' At the same time, the Home Office is developing performance measures that will not be made public but form part of the information available to the inspectorate of police.

Technical and analytical issues

As with many public services, the issue of performance measurement within the police service is subject to a number of technical constraints. As Tilley (1994) points out, measurement issues in this area include: problems measuring the absence of events, wide fluctuations in local crime rates, national changes which impact on crime in economic or policy terms and which may affect crime in ways beyond local control, the fact that crime surveys are generally unable to pick out local patterns of change, and the variability in public reporting and police recording practices which can affect the utility of police data. Collier (2006) highlights the concern – ongoing in all areas of performance monitoring – that the focus tends towards indicators which are easy to measure, or those which encompass a rather traditional view of police work, leaving some significant areas overlooked.

Given the variety of technical issues linked to the diversity of circumstances in which police operate, the initial tendency was for an excess of indicators. This problem was addressed by the Spottiswoode Report on improving police performance monitoring in 2000. According to this study, 'there is a plethora of indicators and information about police outputs and outcomes. But, to date, it has not been possible to draw this information together to build a comprehensive or systematic measure of relative police *efficiency* in meeting their ultimate objectives of promoting safety and reducing crime, disorder and the fear of crime'. The report advocated reducing the number of indicators employed in the Police Performance Assessment Framework and setting differentiated performance targets according to the specific circumstances of different police forces. The study also recommended the joint use of two of the most advanced relative efficiency measuring techniques – Stochastic Frontier Analysis and Data Envelopment Analysis.

> "The localism agenda is likely to create the most significant technical challenges in the future development of police performance monitoring"

These techniques have been used for measuring the relative efficiency of regulated private sector industries, and are increasingly being used in the public sector in other countries. The techniques, however, are controversial. Stone (2002) presents a critique of this methodology and argues that the efficiency measures employed are open to 'gaming': 'to get an efficiency of 100% all a police force has to do is to engineer that it has the uniquely largest value for any of its outputs, since there will be no other force that has the same outputs and thus a smaller input with the same outputs'. He also points out that two forces could have their rankings reversed by an extraneous change in a third force (the principle of third party invariance). Stone suggests that there may be no completely satisfactory way of aggregating performance measures for police forces.

As well as questions surrounding the way overall scores should be presented and compared, there are ongoing issues regarding the types of indicators which best reflect police performance. The idea of crime levels remains a particularly contentious area. There remain two methods of recording crime: the British Crime Survey (which monitors public experience of both reported and unreported crime) and police reporting of Notifiable Offences. Social changes are also altering the standards by

which police performance is now assessed. As Ratcliffe (2002) points out, while crime appears by most measures to have dropped substantially over the last decade, surveys consistently reveal that crime is at, or very near, the forefront of public concern. The strong public interest in, and media spotlight on, crime has produced a situation in which, he argues, fiction and fact have become intertwined, and the public are generally not good at evaluating their realistic likelihood of becoming victims of crime. As such, 'police services are now attempting to counter not just the level of criminal activity, but also the public perception of crime victimisation' (212).

There also remain issues around whether the incidence of crime is a measure of police effectiveness or a measure of demand. The answers to this question largely depend on the extent to which scores should be *contextualised* – a debate which parallels those taking place in schools. The issue of contextualisation of police performance data is explored by Ashby (2005). He argues that the statistics and data structures necessary for comparative performance are underdeveloped in the British public sector, and that current indicator measures on crime are 'crude' and 'may not fully account for the heterogeneity of policing environments and challenges within any one force area' (416). One example is the attempt to compare police forces using 'most similar force' comparisons – aiming to ensure that, say, large urban forces are only compared to other large urban forces. This approach has been criticised, particularly from forces operating in London. He recommends the development of localised targets of specific relevance to the communities being served, with performance evaluated across a national framework founded upon 'neighbourhood types'. The profiling of neighbourhoods has been an area of growing interest – for both commercial and political purposes – particularly via the development of the 'Mosaic' tool in the last few years. Ashby believes these advancements could form the basis of a much more nuanced framework for police performance measurement.

The localism agenda is likely to create the most significant technical challenges in the future development of police performance monitoring. This is an area in which government interest has been growing; a 2004 Home Office White Paper identified a real need to 'make changes to the way police performance is measured and inspected so that it reflects the priorities of the public and their views about the policing they have received' (Ashby, 2005). Finally, as with schools, there is the issue of uncertainty and the requirement that interval estimates are provided

to reflect this, and that detailed comparisons between forces should recognise it.

Usability issues

The questions surrounding the accessibility and usability of police performance indicators have been closely linked to the debate about whom the data are for. The demand for information on police performance from service users has been far less prominent than in the field of education, largely due to the fact that user choice is more restricted. Thus, early studies on police performance data largely focus on its accessibility for police authorities. For example, the Spottiswoode Report on police performance measures in 2000 discusses the problems created by the use of numerous indicators largely in relation to police forces themselves, arguing that they 'lack good measures to fulfil their obligations and to compare their performance' and that the data provided means 'they do not always know what the scope for efficiency gains is or even where they should be looking for them'. Collier (2001) also discusses the problems for police authorities in interpreting the differing counting rules for crime recording, such as the discrepancies between Home Office statistics and the British Crime Survey.

It is only much more recently that the accessibility of police performance data to the wider public has begun to be explored and, as such, this appears to be an under-researched area. Ashby (2005) argues that there remain numerous inconsistencies between current performance assessment frameworks and analytical techniques, and a political agenda which appears to be focusing on fostering more localism and community engagement within the police force.

Ratcliffe (2002) points out that crime mapping is now widely used in the United States to provide the public with more accurate sources of information on crime. This is an area which has also been picked up on in recent Home Office literature (Home Office, 2010). However, he also acknowledges the problems it can engender; issues around personal privacy mean that data must always be aggregated, which creates a further source of error. Additionally, some types of crime remain consistently under-reported, which can distort official data.

In short, it appears that the biggest issue regarding usability will be the way in which the complex variety of data on crime and police perfor-

mance can be presented in a way which both fosters local engagement and offers the public an accurate and comprehensible source of information. All of this argues for careful evaluation of the introduction of new information sources.

Political, ethical and societal issues

As with performance measurement in any other public service, the collection of data on policing has been linked to wider political and societal implications. Again, many of the problems are linked to the diversity of aims and stakeholders attached to public services, and the complexities of their mission. Performance data also needs to address the diversity of activities undertaken by the police, as well as the 'lack of any general agreement as to whether the purposes of the police should be prioritised in favour of crime prevention, crime detection, addressing the public's fear of crime or the wide-ranging non-crime-related services' (Collier, 2006: 166). Collier (2006) also addresses the question of what we consider to be police 'performance': it may be what is done (measured via levels of crime and detection rates); how it is done (public satisfaction with policing); or the results of what is done (the extent to which policing establishes conditions in which the public does not fear crime, the criminal justice system works cooperatively to detect and punish crime and other agencies work to tackle the socio-economic causes of crime).

Collier (2006) reviews the political priorities behind the collection of police performance data and argues that, thus far, 'the development of performance indicators has been primarily top-down with a dominant concern for enhancing control and upwards accountability rather than promoting learning and improvement' (165). He also argues that priorities for the police have been subject to more alterations than for virtually any other public service. The continually shifting focus of performance indicators has prevented forces from investing in the personnel and training which could lead to sustainable long-term improvements in performance. The short-termist nature of many government targets is likely to result in short-term initiatives on the part of the police.

As in education, one of the major concerns from a societal perspective is the potential for perverse incentives or behaviour in a police force dominated by performance monitoring. Scott (1998) picked up on this

problem in a study which took place just after the imposition of performance monitoring in 1995. In interviews with Police Sergeants and Inspectors, he found that they had begun to allocate more resources to tasks associated with performance indicators; for example, on arrest decisions, more 'low-quality' arrests were being made. This trend appears to have continued: Collier (2001, 2006) finds that a culture of performance monitoring meant that the 'quantifiable' dimensions of police work were being prioritised over the 'qualitative'; the speed of response to 999 calls is deemed more significant than the public's satisfaction with the response itself, and the problem of anti-social behaviour – a pervasive concern amongst the public – was in danger of being neglected due to the difficulty of measuring performance in this area. The latest consultation document on policing (Home Office, 2010) also refers to the perverse incentives of a target-centred performance culture. One example is the 'Offences Brought to Justice' target, which has incentivised officers to pursue easy to achieve, low-level detections, rather than focusing on more serious cases.

These individual examples point to an overarching concern, which is echoed in a number of studies in this area. There is a general acceptance that performance management has increased managerial accountability to government, particularly by encouraging police authorities to focus on the costs and accounting needs of the sector (Scott, 1998). However, this increased upward accountability is widely perceived to be encouraging a more 'reactive' policing style that is less engaged with the concerns of the public (Scott, 1997; Ashby, 2005; Collier, 2006). As Ashby (2005) puts it, 'the impending danger of a performance culture is the increased focus upon targets to the detriment of more difficult-to-quantify objectives, the subsequent reversion to more reactive policing styles and the failure to engage with the needs of the local community'. Scott's (1998) interviews with local Police Consultative Committees found that the performance indicators used were often not of importance to local representatives; 'the Audit Commission's view of empowering the consumer of policing services by increasing their knowledge of the costs of policing appeared to be marginal in comparison to requests to the police to deal with localised qualitative problems which were generated by, and causing problems to, specific groups of people (for example, robberies of the elderly and problems with youth behaviour)'. (285)

Collier (2001) argues that the clash between the emphasis on value for money imposed by top-down targets and the demands and priorities of

local people will become an increasingly significant tension. Discussing the impact of the passing of the Human Rights Act on policing, he states that, 'a continuation of the performance culture in an environment where human rights are more important, may lead to tensions and failures in the qualitative aspects of police performance' (38). Scott (1998) also believes that 'the neo-liberal model of the public sector' imposed by performance indicators was shifting the focus away from philosophical ideals of policing by consent (287).

> "The clash between the emphasis on value for money imposed by top-down targets and the demands and priorities of local people will become an increasingly significant tension"

There is some evidence that the perceived 'neo-liberal culture' imposed by public sector performance monitoring is also a matter of concern to police officers themselves. Scott (1998) reports widespread resistance to the competitive element of performance monitoring; 'the majority of officers for both divisions were not motivated by the need to achieve performance levels and they were highly resistant to competing with each other' (286). In 2010, controversy erupted over bonus payments paid to police officers who had achieved their performance targets. Paul McKeever, chairman of the Police Federation, stated that, 'bonuses are being given for the job we should be doing anyway and have not increased productivity. They are also divisive, because they are not received by all officers' (Press Association, 2010).

Collier (2001) puts the case most strongly for a complete re-evaluation of the performance culture in police services. He argues that the prioritisation of quantitative objectives is leading to the 'failure of policing in qualitative terms'. He calls for a 'values-based learning paradigm', which does not abandon the drive for efficiency fostered by performance monitoring but attempts to reconcile it with the challenges of day-to-day policing; 'such a paradigm is based on values of public service, integrity and justice that are already found within the police service but which may, in the present control-dominated regime, become increasingly subservient to quantitative measurement' (38).

Issues for further research

The localism agenda is likely to be the main driver of reform to police performance monitoring. The perceived need to engage with the priorities of local communities and to foster improved public accountability within the police force is a recurring feature of both academic and policy discussions. A general conclusion is that, if public satisfaction with policing is to improve, a fuller public discussion of its objectives will be required (Collier, 2006; Ashby, 2005; Scott, 1998). Alongside this new orientation towards communities is the idea that police performance data needs to be better contextualised if it is to provide information that is both accurate and useful.

The challenge for future research will be in identifying the ways in which these aims can be realised in practice. There have been a number of recent proposals regarding the provision of information on policing to a wider public, including crime mapping and increasing the availability of data. Interestingly, given the limited nature of user choice within police services, comparative league tables are unlikely to be a prominent feature of these developments. However, this leaves open the question of how such data can be presented in an accurate and accessible way, negotiating the ongoing trade-off between complexity and usability. An emphasis on the priorities of communities and the wider public will also fuel debates on the less quantifiable aspects of policing, which are currently neglected by performance measures. The incorporation of qualitative, alongside quantitative, measures will most likely prove a key challenge for future performance frameworks.

Summary of issues and recommendations

We have highlighted a variety of technical, political and social issues linked to the growth of league tables in policing. A summary of these issues and implied recommendations to address them are below:

- Current indicator measures on crime do not fully account for the heterogeneity of policing environments and challenges within different force areas. Crime should be contextualised in terms of local conditions.
- Problems with the recording of crime, especially where high stakes targets are in place, need to be addressed.
- The uncertainty attached to statistical estimates, especially for small areas, should be addressed.

- Increased accountability to the government is perceived to be encouraging a more 'reactive' policing style that is less engaged with the local community. Locally defined outcomes should be incorporated.
- It is only recently that the accessibility of police performance data to the wider public has begun to be explored. There should be better information for public understanding of the complex variety of data.
- As in education, one of the major concerns is the potential for perverse incentives or behaviour in a police force dominated by performance monitoring. There should be thorough evaluation of side effects and perverse incentives.
- There is some evidence that the perceived 'neo-liberal culture' imposed by public sector performance monitoring is also a matter of concern to police officers themselves. The relevance of the idea of 'competition' in a market sense among police forces should be examined.
- The role of unpublished rankings available to the inspectorate for discussion with individual police forces should be explored.

4 General conclusions and recommendations

In discussing the three areas of schools, higher education and policing this report has raised a number of issues for further debate. These can be summarised as follows:

Rankings as a tool for improvement

An issue that pervades the discussion in this report is how far a culture centred on the publication of performance data detracts from seeking an understanding of *how* to improve education or policing. Critics raise concerns that it is a 'sideshow' that appears to provide incentives to improvement while failing to evaluate the reasons for success.

In school education, much research is devoted to finding what works and why. The data used to compile league tables, while limited in terms of what is measured, is large and comprehensive and is being used by researchers to help understand the system, for example by looking at neighbourhood versus school effects (see www.bris.ac.uk/cmpo/plug). This is an encouraging development.

A particularly important issue is where comparative rankings are used as 'screening' devices that are not published but used as part of an institutional improvement programme. In principle this could obviate some of the currently perceived negative effects of league tables, although it would require a level of trust in professional judgments that may be difficult to achieve. Nevertheless, it seems that this is a worthwhile aim and should be given serious attention.

Who should produce comparative rankings?

In the UK, apart from in higher education, it is effectively government departments that facilitate the collection of performance measures. Yet this need not be the case. In higher education for example, the *THE* in the UK, the work of the Carnegie Commission in the US and the example of Germany suggest that it is possible for independent institutions to undertake this task. Germany in particular has demonstrated that a balanced presentation of results is possible.

If performance indicators are to continue, then serious consideration should be given to the delegation of their production to independent institutions with a wide representation of advisors and suitable technical support, as in the case of Germany (see p32).

Technicalities

Where performance data are published their technical limitations should be highlighted. This will include factors such as the choice of measures, aggregation and uncertainty resulting from relatively small numbers. Such technical issues are central to the provision of validity of rankings and further research on how these can be presented is important.

Monitoring

One of the deficiencies at the present time is the paucity of satisfactory evaluations of the effects of performance data on institutional performance. More research is needed, especially where new league tables are proposed. Pilot studies that are properly evaluated are one useful means of carrying this out.

It is clear that performance monitoring and targets can act as powerful incentives for behavioural change amongst public service providers. On the one hand, this means they can potentially serve as an important driver of organisational improvements. Yet, where targets are badly chosen or poorly specified, there is a danger they can in fact undermine other desirable outcomes. Furthermore, without awareness of their inevitable limitations there is a danger that policymakers, professionals and the general public come to place too great an emphasis on their outcomes. The future challenge for UK policymaking is to accept the

need to modify the present regime so that the positive aspects can be emphasised and the negative ones diminished.

Finally, it is worth raising the question of whether an independent body could have a role in monitoring developments, providing guidelines for good practice and also becoming involved in the production and presentation of performance indicators.

Summary of recommendations

General

- Serious consideration should be given to using comparative rankings as 'screening' devices that are not published or made available beyond those institutions involved, but used as part of an institutional improvement programme, so that institutions can seek improvement without perverse incentives arising from full public exposure. We refer to this as 'intelligent accountability'. This could obviate some of the currently perceived negative effects of league tables.
- Wherever league tables are published they should be accompanied with appropriate and prominent 'health warnings' highlighting their technical limitations. These should include assessments of the statistical uncertainty, often large, that may limit their usefulness. They should also include statements about the quality of the measurements that go to make up the indicators, including the effects of aggregation. In a broader context, there is a need for a debate about whether simply making data available to citizens will encourage good use of them. In the absence of professional support and advice, data analysis can be very difficult for those with limited experience or expertise. Deliberate or unintentional misuse of statistical information should not be encouraged and there is a real danger that this could occur increasingly unless public awareness of the issues improves.
- More research is needed on the effects of performance data on institutional performance. There should be careful evaluation of existing league table systems and the systematic piloting of proposed systems. This evidence should pay particular attention to 'knock-on' effects whereby resources may be reduced for some important activities in order to improve league table performance.
- Consideration should be given to whether one or more independent (not for profit) institutions could have a role in monitoring develop-

ments, providing guidelines for good practice and also become involved in the production and presentation of performance indicators. Such institutions should be independent of government.

Education

- The linking of league tables to rewards should be weakened to reduce the side effects of inappropriate 'gaming' and to reduce stress among teachers, parents and students. This would also have the desirable effect of making the results a more objective evaluation of performance. The problematic consequences for schools serving the most disadvantaged pupils particularly need to be addressed.
- The government should consider ways to prevent league tables being exploited by the media, such as ensuring that measures of uncertainty are provided around any institutional results. Associated with this there could be a campaign to better inform the public at large about the strengths and limitations of league tables, although any such attempt poses considerable challenges.
- Consideration should be given to alternative ways of using quantitative information to monitor educational performance generally. This can be achieved by in-depth study of a sample of schools and students within a national database. A useful model is the Assessment of Performance Unit that was set up in the 1970s in England and discontinued in the 1980s (Gipps and Goldstein, 1983).
- Consideration should be given to using performance information as a screening device rather than publishing as league tables, as in the Hampshire experiment. This could be accompanied by an emphasis on evaluation and inspection systems that are designed to emphasise ways of assisting schools to cope with problems rather than 'exposing' them using public rankings.
- Ways to rely less on a small number of indicators should be sought, as well as those which cover more aspects of learning.
- More appropriate statistical analysis models should be used to describe institutional differences that allow for differential performance for different groups of students. In particular, there should be a shift away from the comparison of individual institutions towards research that helps to identify modifiable factors that appear to be related to good performance.
- An ethical code to govern the publication of school performance measures should be formulated, as suggested by Goldstein and Myers (1996). This would be based on two broad principles: that

unjustified harm to those to whom the information applies should be prevented, and that there should be no *absolute* publication rights for performance data.
- Further consideration needs to be given to the role of inspection and accreditation agencies as a means of evaluating individual institutions. Trust in such agencies may not be easy to achieve, especially when they are perceived to be instruments of government. A discussion of such agencies is given in Appendix A.

Higher education

- Indicators need to be selected according to validity rather than availability as currently tends to be the case. This implies more qualitative and process indicators, although care needs to be exercised in terms of their subjectivity.
- Disaggregated indicators are important and the temptation to aggregate into one index, or even a small number of indexes, should be resisted.
- Measures of uncertainty need to be displayed.
- For users, broad categories rather than precise rankings are to be preferred and sensitivity analyses with different weightings to components should be conducted to test stability.
- Subject-based rankings should be emphasised.
- Further consideration needs to be given to the role of inspection and accreditation agencies as a means of evaluating individual institutions. Trust in such agencies may not be easy to achieve, especially when they are perceived to be instruments of government. A discussion of such agencies is given in Appendix A.

Policing

- Current indicator measures on crime do not fully account for the heterogeneity of policing environments and challenges within different force areas. Crime should be contextualised in terms of local conditions.
- Problems with the recording of crime, especially where high stakes targets are in place, need to be addressed.
- The uncertainty attached to statistical estimates, especially for small areas, should be addressed.
- Increased accountability to the government is perceived to be encouraging a more 'reactive' policing style that is less engaged with

the local community. Locally defined outcomes should be incorporated.

- It is only recently that the accessibility of police performance data to the wider public has begun to be explored. There should be better information for public understanding of the complex variety of data.
- As in education, one of the major concerns is the potential for perverse incentives or behaviour in a police force dominated by performance monitoring. There should be thorough evaluation of side effects and perverse incentives.
- There is some evidence that the perceived 'neo-liberal culture' imposed by public sector performance monitoring is also a matter of concern to police officers themselves. The relevance of the idea of 'competition' in a market sense among police forces should be examined.
- The role of unpublished rankings available to the inspectorate for discussion with individual police forces should be explored.

References

Ashby, D. I. (2005) 'Policing Neighbourhoods: Exploring the geographies of crime, policing and performance assessment', in *Policing and Society* 15(4).

Baty, P. (2010) 'Measured, and found wanting more' in *Times Higher Education,* 8th July 2010.

Berry, C. (1999) 'University league tables: artefacts and inconsistencies in individual rankings' in *Higher Education Review,* 21(2).

Bevan, G. and Hamblin, R. (2009), 'Hitting and missing targets by ambulance services for emergency calls: effects of different systems of performance measurement within the UK', in *Journal of the Royal Statistical Society,* Series A (Statistics in Society), 172: 161–190.

Bird, S. *et al*. (2005), 'Performance indicators: good bad, and ugly' in *Journal of the Royal Statistical Society*, A. 168: 1–27.

Bowden, R. (2000) 'Fantasy Higher Education: University and college league tables' in *Quality in Higher Education,* 6(1).

Burgess, S., Propper, C. and Wilson, D. (2002), 'Does Performance Monitoring Work? A Review of the Evidence from the UK Public Sector, Excluding Health Care', CMPO Working Paper Series No. 02/49.

Cameron, D. *Giving Power Back to the People,* Speech to Imperial College London, 25th June 2009.

Clarke, M. (2002), 'Some Guidelines for Academic Quality Rankings' in *Higher Education in Europe,* 27(4).

Collier, P. M. (2001) 'Police Performance Measurement and Human Rights' in *Public Money and Management,* 21(3).

Collier, P. M. (2006), 'In Search of Purpose and Priorities: Police performance indicators in England and Wales' in *Public Money and Management,* 26(3).

Dill, D. and Soo, M. (2005) 'Academic Quality, League Tables and Public Policy: A cross-national analysis of university ranking systems' in *Higher Education*, 49(4).

Drake, L. and Simper, R. (2003), 'The Measurement of English and Welsh Police Force Efficiency: A comparison of distance function models' in *European Journal of Operational Research*, 147(1).

Eccles, C. 'The Use of University Rankings in the United Kingdom' in *Higher Education in Europe,* 27(4).

Federkeil, G. (2002), 'Some Aspects of Ranking Methodology: The CHE rankings of German universities' in *Higher Education in Europe,* 27(4).

Gipps, C. and Goldstein, H. (1983), *Monitoring Children* (London, Heinemann).

Goldstein, H. and Spiegelhalter, D. J. (1996), 'League tables and their limitations: statistical issues in comparisons of institutional performance' in *Journal of the Royal Statistical Society,* A. 159: 385–443

Goldstein, H. and Leckie, G. (2008) 'School League Tables: What can they really tell us?' in *Significance,* 5(2).

Goldstein, H. and Myers, K. (1996), 'Freedom of Information: Towards a code of ethics for performance indicators' in *Research Intelligence,* 57.

Goldstein, H., Burgess, S. and McConnell, B. (2007), 'Modelling the impact of pupil mobility on school differences in educational achievement' in *J. Royal Statistical Society,* A. 170: 941–954.

Gormley, W. R. and Weimer, D. L. (1999), *Organisational Report Cards, (*Cambridge, Mass: Harvard University Press).

Hallgarten, J. (2001) 'School League Tables: Have they outlived their usefulness?' in *New Economy* 8(4).

HEFCE, (2008) 'Counting What is Measured or Measuring What Counts? League tables and their impact on higher education in England', Issues Paper 14.

Home Office (2010), 'Policing in the Twenty-First Century: Reconnecting police and the people', (London: HMSO).

Hood, C. (2007), 'Public Service Management by Numbers: Why does it vary? Where has it come from? What are the gaps and puzzles?' in *Public Money and Management*, Vol. 27, No.2.

Kahneman, D. and Tversky, A. (1971), 'Belief in the law of small numbers' in *Psychological Bulletin,* 76(2).

Keasey, K., Moon, P. and Duxbury, D. (2000), 'Performance measurement and the use of league tables: Some experimental evidence of dysfunctional consequences' in *Accounting and Business Research*, 30(4).

Ladd, H. F. and Walsh, R. P. (2002), 'Implementing Value-Added Measures of School Effectiveness: Getting the incentives right' in *Economics of Education Review,* 21(1).

Leckie, G. and Goldstein, H. (2011), 'Understanding uncertainty in school league tables', in *Fiscal Studies* (Forthcoming).

Leckie, G. and Goldstein, H. (2009), 'The limitations of using school league tables to inform school choice' in *Journal of the Royal Statistical Society,* A 172: 835–851.

Mansell, W. (2007), *Education by numbers* (London, Politico).

Newton, X., Darling, A., Hammond, L., Haertel, E. and Thomas, E. (2010), 'Value-added modelling of teacher effectiveness: an exploration of stability across models and contexts', Educational Policy Analysis Archives, 18, 23.

Oswald, A. (2002), 'An Economist's View of University League Tables' in *Public Money and Management,* 21(3).

Press Association (2010), 'Police bonus payments £150m a year', 14th August 2010.

Propper, C. and Wilson, D. (2003), 'The Use and Usefulness of Performance Measures in the Public Sector' in *Oxford Review of Economic Policy,* 19(2).

Ratcliffe, J. H. (2002), 'Damned if you don't, damned if you do: Crime mapping and its implications in the real world' in *Policing and Society*, 12(3).

Roberts, D. and Thompson, L. (2007), 'University League Tables and the Impact on Student Recruitment', *Reputation Management for Universities*, Working Paper Series No. 2, The Knowledge Partnership.

Robinson, P. (2001), 'IPPR Indicators: Wage premiums and league tables' in *New Economy*, 8(4).

Russell, B. (2001), 'Northern Ireland abandons league tables for schools' in *The Independent*, 11th January 2001.

Scott, J. (1998), 'Performance Culture: The return of reactive policing' in *Policing and Society*, 8(3).

Smith, P. (1995), 'On the unintended consequences of publishing performance data in the public sector' in *International Journal of Public Administration,* 18.

Spottiswoode, C. (2000), 'Improving Police Performance: A new approach to measuring police efficiency', Public Services Productivity Panel.

Stone M. (2002), Public Money www.informaworld.com/smpp/title~db=all~content =t793706091" *Management*, 22, 33–40.

Terenzini, P. T. and Pascarella, E. T. 'Living with Myths: Undergraduate education in America' in *Change*, 26(1).

Thanassoulis, E. (1995), 'Assessing Police Forces in England and Wales using Data Envelope Analysis' in *European Journal of Operational Research,* 87.

Tilley, N. (1994), 'Thinking about Crime Prevention Performance Indicators', Crime detection and Prevention Series Paper 57.

Turner, D. (2005), 'Benchmarking in Universities: League tables revisited' in *Oxford Review of Education,* 31(3).

Vaughn, J. (2002), 'Accreditation, Commercial Rankings, and New Approaches to Assessing the Quality of University Research and Education Programmes in the United States' in *Higher Education in Europe,* 27(4).

Visscher, A. J. (2001), 'Public School Performance Indicators: Problems and recommendations' in *Studies in Educational Evaluation,* 27 (3).

Wagner, L. (1998), 'What is the Point of University League Tables?' in *CVCP News* (July 1998).

Wiggins, A. and Tymms, P. (2002), 'Dysfunctional Effects of League Tables: A comparison between English and Scottish primary schools' in *Public Money and Management,* 22 (1).

Wilson, D. (2004), 'Which Ranking? The impact of a 'value-added' measure on secondary school performance' in *Public Money and Management*, 24(1).

Wilson, D., Croxson, B. and Atkinson, A. (2006), 'What Gets Measured Gets Done' in *Policy Studies,* 27(2).

Yorke, M. (1998), '*The Times* league table of universities, 1997: A statistical appraisal' in *Quality Assurance in Education*, 6(1).

Appendix A: Alternatives to league tables

A: Inspection agencies

Inspection agencies are publicly funded bodies designed to provide 'in-depth' evaluations of institutional functioning using the judgments of experienced professionals. Here we have Ofsted at school-level and to some extent the HEFCE research assessment exercise at higher education (HE) level and its successor the Research Excellence Framework (REF). We have HMCIC for the police force. The QAA is formally an accreditation agency (see below).

Schools and non-HE education: Ofsted

Ofsted regulates and inspects childcare and children's social care and inspects schools, colleges, initial teacher education, work-based learning and skills training, adult and community learning, education and training in prisons and other secure establishments and the Children and Family Court Advisory Support Service (Cafcass). It assesses children's services in local areas, and inspects services for looked-after children and child protection. It seeks to promote improvement in services inspected and regulated, and to ensure that they focus on the interests of the children and young people, parents and carers, adult learners and employers who use them. It is also concerned about providing value for money.

For further information about Ofsted visit:
www.ofsted.gov.uk/Ofsted-home/About-us

Higher education

The Research Assessment Exercise (RAE), and its successor the Research Excellence Framework (REF) in the UK, are examples of academic members of universities, together with a small number of non-academics, assessing the research quality of colleagues. The RAE

has been largely qualitative but with increasing demands for the use of quantitative information in the form of journal and author citations to be used in the REF. The debate around this issue has focused on the limitations of quantitative measures: their inability to capture important aspects of research quality; their inability to take a long-term perspective; and their misleading nature that derives from the manner of their construction, including coverage. The most important advantage of quantitative measures appears to be their relative cheapness.

Police force

A joint framework for inspections was approved in 2009 allowing for a joint programme with the Audit Commission, although with the dismemberment of the latter it is not clear how this will operate in the future. The following is a summary of the role of the inspections.

Her Majesty's Inspectors of Constabulary are appointed by the Crown on the recommendation of the Home Secretary and report to Her Majesty's Chief Inspector of Constabulary, who is the Home Secretary's principal professional policing adviser. Her Majesty's Inspectors of Constabulary are charged with examining and improving the efficiency of the Police Service in England, Wales and Northern Ireland. HMIC is independent both of the Home Office and of the Police Service.

The primary functions of HMIC include:

- The formal inspection and assessment of all forces in England, Wales and Northern Ireland (as well as a number of non-Home Office funded police forces), HM Revenue and Customs and the Serious Organised Crime Agency.
- Undertaking thematic inspections across forces, some in conjunction with other bodies, including the other Criminal Justice System Inspectorates.
- Undertaking a key advisory role within the tripartite system (Home Office, chief officer and police authority/Northern Ireland Policing Board), where its independence and professional expertise are recognised by all parties. HMIs also provide a crucial link between forces and the Home Office, and contribute to the process of appointments to the most senior ranks in the Police Service.

For further information about HMIC, visit:
www.inspectorates.homeoffice.gov.uk/hmic

Issues

The following issues arise:

- The inspections are costly.
- For any given institution an inspection may only occur infrequently so that information will be out of date.
- The amount of material produced may be difficult to absorb and summaries may be perceived to be potentially misleading.
- In some cases inspection teams may be influenced unduly by quantitative measures used in league tables. Ofsted school inspectors for example, are expected to take account of these.
- Relevant but sensitive information about individuals may be released in ways that breach individual human rights.
- Is it appropriate that some (or even all) parts of such reports should be considered as a private document whose purpose is to raise issues that a governmental funding body can discuss with an institution, rather than as a public accountability instrument or one that is designed to inform users? This was largely the traditional function of school inspections. More generally, how much should be in the public domain and how much reserved for private consultation?
- How can inspection agencies remain independent of political pressures from government?

B: Accreditation agencies

Typically these are formed by a group of institutions that agree to submit themselves to a process of inspection and reporting designed to establish and maintain quality standards.

QAA

The primary responsibility for academic standards and quality in UK higher education rests with individual universities and colleges, each of which is independent and self-governing. QAA checks how well they meet their responsibilities. It seeks to identify good practice and make recommendations for improvement. It also publishes guidelines intended to help institutions develop effective systems to ensure students have high quality experiences. All of its institutional reports are available from its website.

It carries out the following activities:

- Conducting reviews of universities and colleges
- Publishing reports on the confidence that can be placed in an institution's management of standards and quality
- Providing guidance to universities and colleges on maintaining academic standards and improving quality, in line with the Academic Infrastructure
- Investigating causes for concern about academic standards and quality
- Advising governments on applications for degree awarding powers and university title
- Engaging with European and wider international developments

QAA is an independent body funded by subscriptions from universities and colleges and through contracts with the higher education funding bodies. It carries out external quality assurance by visiting universities and colleges to review how well they are fulfilling their responsibilities. As such it has some of the characteristics of an accreditation agency.

For further information about QAA visit:
www.qaa.ac.uk/aboutus/WhatWeDo.asp

Open & Distance Learning Quality Council
For details visit: www.odlqc.org.uk/

ODL QC was founded in 1969 as the Council for the Accreditation of Correspondence Colleges, becoming the Open and Distance Learning Quality Council in 1995. Set up at the request of government, it continues to have governmental support and cooperation, though it is now an independent body, and a registered charity.

The aim of the Council is to identify and enhance quality in education and training for open and distance learning, and to protect the interests of learners.

The Council sets out definitions of quality and standards and open and distance learning providers that meet those standards are eligible to apply for accreditation by the Council.

Accreditation follows an assessment of a provider's administrative and tutorial methods, educational materials and publicity, to ensure that all standards are met. Once accredited, providers are monitored to ensure that students continue to receive good service, and are re-assessed at least once every three years.

The Council includes representatives of accredited providers as well as members drawn from professional and public bodies involved in education, chosen for their ability to contribute to the work of the Council and all highly qualified in their own particular fields.

Business schools

There are a few specialised agencies set up for business schools. These are as follows:

- The Association of MBAs (www.mbaworld.com)
- The Foundation for International Business Administration Accreditation (www.enqa.eu/agencydet.lasso?id=22)
- The Accreditation Council for Business Schools and Programs (www.acbsp.org/p/st/ld/sid=s1_001)
- The Association to Advance Collegiate Schools of Business (www.aacsb.edu/)
- The International Assembly for Collegiate Business Education (www.iacbe.org/)

Issues

- Since the institutions fund the accreditation body, there is a moral hazard that derives from a pressure to provide positive ratings.
- In theory a 'market' in possible bodies can ensure maintenance of standards through a mechanism of user confidence and choice, but in HE this is unlikely to operate satisfactorily since there are too few agencies in the field.
- Such 'markets' have not worked in similar situations such as that of credit rating agencies.
- Are there alternative public agencies that could take over the role of QAA, like Ofsted, providing more independent monitoring?
- How transparent can accreditation agencies be about their procedures?
- How can conflicts of interest be resolved?

Appendix B: Forum participants

Participants in a policy forum on league tables in the public sector, held by the British Academy Policy Centre on 19 January 2011:

- Stephen Ball FBA, Institute of Education
- David Bartholomew FBA, London School of Economics and Political Science
- Phil Baty, Times Higher Education
- Paul Black, Kings College London
- Selina Chen, British Academy
- Rob Copeland, University and College Union
- Colin Crouch FBA, University of Warwick
- Stephen Crump, University of Newcastle
- Beth Foley, British Academy
- Howard Glennerster FBA, London School of Economics and Political Science
- Harvey Goldstein FBA, University of Bristol
- Yvonne Hawkins, Higher Education Funding Council for England
- Ellen Hazelkorn, Dublin Institute of Technology
- Christopher Hood FBA, University of Oxford
- Siôn Humphreys, National Association of Head Teachers
- Vivienne Hurley, British Academy
- John Kirkpatrick, Audit Commission
- George Leckie, University of Bristol
- Jovan Luzajic, Universities UK
- Amobi Modu, Home Office
- Mike Pidd, Lancaster University Management School
- Bernard Silverman, Home Office
- Deborah Wilson, University of Bristol

More details on the policy forum can be found online at www.britac.ac.uk/policy/League-tables-in-the-public-sector.cfm

British Academy Policy Centre publications

Raising household saving, a report prepared by the Institute for Fiscal Studies for the British Academy, February 2012

Post-immigration 'difference' and integration: The case of Muslims in western Europe, a report for the British Academy project *New paradigms in public policy,* February 2012

Building a new politics?, a report for the British Academy project *New paradigms in public policy,* January 2012

Squaring the public policy circle: Managing a mismatch between demands and resources, a report for the British Academy project *New paradigms in public policy,* November 2011

Economic futures, a report for the British Academy project *New paradigms in public policy,* September 2011

Climate change and public policy futures, a report for the British Academy project *New paradigms in public policy*, July 2011

History for the taking? Perspectives on material heritage, a British Academy report, May 2011

Stress at work, a British Academy report, October 2010

Happy families? History and family policy, a British Academy report, October 2010

Drawing a new constituency map for the United Kingdom: The parliamentary voting system and constituencies bill 2010, a British Academy report, September 2010

Choosing an electoral system, a British Academy report, March 2010

Social science and family policies, a British Academy report, February 2010

Punching our weight: The humanities and social sciences in public policy making, a British Academy report, September 2008

The British Academy, established by Royal Charter in 1902, champions and supports the humanities and social sciences across the UK and internationally. As a Fellowship of 900 UK humanities scholars and social scientists, elected for their distinction in research, the Academy is an independent and self-governing organisation, in receipt of public funding. Its Policy Centre, which draws on funding from ESRC and AHRC, oversees a programme of activity, engaging the expertise within the humanities and social sciences to shed light on policy issues, and commissioning experts to draw up reports to help improve understanding of issues of topical concern. This report has been peer reviewed to ensure its academic quality. Views expressed in it are those of the author(s) and are not necessarily endorsed by the British Academy but are commended as contributing to public debate.

Institutional rankings or 'league tables' are now widely used in the public sector. Employed in areas such as health, policing and education they help determine whether schools or hospitals are deemed to be 'failing', whether police forces are tackling crime effectively and how students rate their university courses. But despite being widespread, their application is still highly contentious.

Measuring Success examines the use of league tables in education and policing, and reviews the available evidence to determine the benefits and the problems associated with their use. The report concludes that good evidence about league tables is in short supply, which has only helped fuel their controversy; it also highlights the limitations of league tables and recommends that wherever they are produced they should be accompanied by prominent 'health warnings'. Furthermore, the authors suggest that some of the negative side effects of league tables could be reduced if they are used only as an internal tool to improve performance by the institutions involved and not published or made publicly available.

Sponsored by

Arts & Humanities Research Council

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

ISBN 978-0-85672-600-2

THE BRITISH ACADEMY
10–11 Carlton House Terrace
London SW1Y 5AH
+44 (0)207 969 5200

Registered Charity: Number 233176

www.britac.ac.uk

BRITISH ACADEMY

POLICY CENTRE