Policy Discussion Briefing

January 2007

Centre For Health Economics

Composite performance measures in the public sector

Rowena Jacobs, Maria Goddard and Peter C. Smith

Introduction

It is rare to open a newspaper or read a government report without seeing some sort of reference to the performance of public sector organisations. The public have become used to judging schools, hospitals, local councils, and other public sector organisations, in terms of their performance rating. This has led to an explosion in the number and variety of league tables. **League table position can have major implications**, for example wholesale changes in leadership of an organisation. As a result, achieving good ratings is a major endeavour for public sector managers.

Although a variety of performance measures exist, current government policy in England emphasises the creation of **composite indicators** (aggregate indicators) in the public sector and they are **used widely** in health, social services, education, local government and other service areas.[1] Whether it is the 'star ratings' of social service departments or the 'research assessment exercise' ratings of universities, the use of a single score that summarises a wealth of underlying performance data, has wide appeal. These composite performance ratings have taken on great importance as they are often used to **reward or penalise organisations**.

Despite the apparent simplicity of composite measures, their use and interpretation raises several challenges.

We examine whether composite indicators are a good way of measuring performance in the public sector. In particular, we want to assess whether such measures are robust (rigorous) and can accurately reflect genuine differences in performance. **We ask**:

- Does the method of constructing composite indicators influence the ratings of organisations?
- Are ratings robust or are they subject to instability?
- Are composite indicators influenced by uncertainty?
- Is some of the variation in performance due to factors beyond the control of managers?
- Are ratings stable over time?

The purpose of this briefing is to summarise the answers to these questions. In brief, **we find**:

- The method of constructing composite indicators matters – different ways of aggregating underlying performance data can have a huge impact on the ratings
- There is a great degree of uncertainty in composite scores – they are not robust
- If we take account of random variation or measurement error in performance indicators, we can estimate genuine performance variations between organisations – these composite scores have much greater precision, and reflect performance which is within managerial control
- Composites are unstable when methods change year on year

This policy briefing describes how we obtained our results and sets out the policy implications.

The Public Services Programme
Quality, Performance and Delivery
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Author details: Rowena Jacobs*, Maria Goddard, Peter C. Smith, Centre for Health Economics, University of York, UK. *rj3@york.ac.uk http://www.york.ac.uk/inst/che/

Table 1: Arguments for and against the use of composite performance measures

Arguments for	Arguments against
Places performance at the centre of the policy arena	By aggregating measures, composites may disguise failings in parts of the system
Offers policymakers a summary of complex multi-dimensional issues	Difficult to determine the source of poor performance and where to focus remedial action
Presents the 'big' picture and may be easier to interpret	A comprehensive composite may rely on poor quality data in some dimensions
Can offer a rounded assessment of performance	If certain performance measures are excluded, it may distort behaviour in undesirable ways
May promote accountability and facilitate communication with the public	Individual performance measures used in composites are often contentious
Easier to track progress of a single indicator over time rather than a whole package of indicators	Composite measures depend crucially on weights attached to each performance dimension, but the methodology for obtaining and using weights is not straightforward
May stimulate better analytic methods and data quality	
Indicates which organisations represent beacons of best performance	
Indicates priority organisations for improvement efforts	

Why use composite performance measures?

Composite indicators provide a **single score** or rating that is **easy to understand** and offer an attractive way of summarising a wealth of performance data, but they also have **several drawbacks**.[2] These are summarised in Table 1.

How are composite performance measures constructed?

Although appearing simple, the **process of creating a composite** from a wealth of disparate performance data is **very complex**[3] and involves a series of judgements at each stage of construction. The steps include:

- choosing the organisations to be assessed
- choosing the organisational objectives to be encompassed
- choosing the indicators to be included
- transforming the measured performance on individual indicators so that they can be aggregated
- combining the individual indicators using addition or other decision rules
- specifying an appropriate set of weights for aggregation purposes
- adjusting for environmental or other uncontrollable influences on performance
- using sensitivity analysis to test the robustness of the composite score

The judgements made at each step can have a profound impact on the composite scores, and hence on the incentives faced by the organisations.

Are composite performance measures different to other performance measures?

Little is known about the degree to which composite measures are an appropriate metric for evaluating performance in the public sector. Many of the research questions we ask are common in considering any type of performance indicator. However, composite performance measures are associated with additional **methodological challenges** that influence the degree to which they may represent an adequate performance measure.

What data do we use?

We use longitudinal data on composite performance measures from two key public services in England, namely the **healthcare** and **local government** sectors. The data comprise:

- NHS Trust star ratings for around 180 acute NHS Trusts from 2000/01 to 2004/05 covering around 40 Performance Indicators.
- Comprehensive Performance Assessments (CPA) ratings for around 150 local authorities (councils) from 2001/02 to 2004/05 covering around 110 Best Value Performance Indicators.

What are our methods?

We **construct a generic composite** indicator for each of the two sectors in order to examine the robustness of composites to the methodological choices involved at each step in the construction.

We choose a subset of 10 performance indicators for healthcare and a subset of 35 performance indicators for local government. Indicators for healthcare include performance measures on patient outcomes, access, satisfaction surveys, data quality and staffing. Indicators for local government cover several key areas such as corporate health, education, social services, housing, benefits, environment, transport, planning, culture and libraries, and community safety.

All indicators are transformed to "more is better" e.g. death rates are converted to survival rates. We then **standardise** the performance indicators, so that they can be aggregated.

In the first instance, to create a composite score for each hospital and local authority, we just add the indicators up giving them an equal weight and **create a ranking** based on the composite score. We have a final sample of 117 hospitals and 97 local authorities.

We explore four main themes:

1. Examining uncertainty

In order to explore the degree of uncertainty on the composite measure, we use simulation methods - these are used to imitate a real-life system.[4] Without the aid of simulation, we would only be able to construct a single composite index for each hospital and local authority. However, we perform 1000 **simulations** which enables us to calculate the range of composite scores a single organisation obtains over 1000 repetitions and thus estimate the degree of uncertainty in the composite.

2. Examining 'real' performance variation

Every performance signal will consist of real variations in performance, variations due to differing local circumstances, and random fluctuations (natural statistical variation). It is important to disentangle different sources of variation on performance measures because we are interested in isolating the **real performance variation** that is **within managerial control**. For each of the performance indicators, we estimate the proportion of variation caused by factors such as measurement error and random fluctuations. The remainder of the variation is assumed to reflect genuine variations in levels of performance.

We estimate the proportion of **random variation** on each indicator. In Figure 1, the estimate of random variation on survival rates in hospitals is 27%. We therefore shrink the variation around each indicator to reflect the estimate of random variation. This is done for each individual indicator. This variation is therefore beyond management control.

3. Examining alternative aggregation methods

In practice, a range of different aggregation methods is often used to construct composites. Different weighting structures can be applied to reflect different priorities attached to attainment on certain indicators. There is little consensus on what weights should be applied to the underlying indicators, or indeed whose preferences these weights should reflect. Decision rules or algorithms are also often applied as an aggregation method to ensure attainment of minimum standards on some indicators, thus introducing indirect implicit weights. Decision rules may, for example, assign *x* score if the organisation is rated good on 3 out of 4 indicators; or *y* if rated good on only 2 out of 4 indicators.

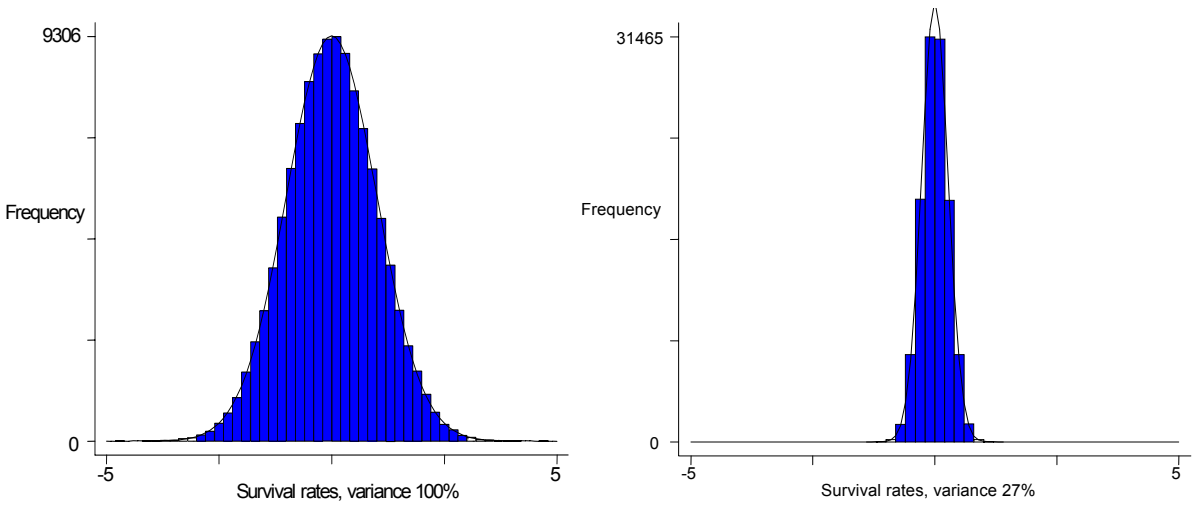


Figure 1: Example of shrinking variation around survival rates from 100% to 27% (estimate of random variation)

We test a range of alternative approaches to aggregating performance indicators into a composite:

- Adding the 10 indicators (healthcare) or 35 indicators (local government) with equal weights (base case);
- Adding the indicators but **varying** the **weights**;
- Applying **decision rules** to assign hospitals or local authorities to ordinal categories (e.g. 0-3 stars or 'excellent', 'good', 'poor'). These types of sequential rules are commonly applied in aggregation methods (e.g. in the Star ratings and CPA).

4. Examining composite indicators over time

We observe a lot of year-to-year changes in performance ratings. Some of this variation may be due to changes in the methodology applied (for example the aggregation methods, weighting systems, and decision rules). We explore how much of the **year-to-year change in performance ratings** is caused by data fluctuations, rather than methodology. We replicate the process of generating a generic composite indicator for each sector for each year, assuming the methodology to construct the composites is constant over time.

What are our findings?

We present a selective summary of our results from healthcare and local government in turn.

1. Examining uncertainty

Our results produce a set of rankings for hospitals and local authorities according to their composite score with a 95% confidence interval reflecting uncertainty around this composite score. These results are illustrated in the left panel of Figure 2. The dark dots show the composite score for hospitals arranged in order from worst to best – the dots assume all variation is due to genuine differences in performance. (This is how most composite indicators are presented in practice). Around each of the dark dots the vertical line shows the 95% confidence interval arising from the simulations – the vertical lines assume all variation is random. This naïve view of variation reflects a **considerable degree of uncertainty** in the composite score since the confidence intervals overlap over almost the entire range of performance. This means we cannot be certain that hospitals with the best composite score are necessarily performing better than those with a middle-of-the-range composite score since a hospital's score could fall anywhere along this vertical line. Similar results are obtained for local authorities.

However, after taking account of **random variation** in the underlying performance indicators (shrinking the variation down to only that element which is beyond managerial control), the results in the right panel of Figure 2 illustrate that we are able to estimate **genuine performance variations**. It is now possible to say with certainty that (for example) the hospitals at the bottom are scoring less well than those at the top, though there is still overlap in the middle. This is clearly a great improvement over presenting composite indicators as just the dark dots (how it is usually done), where we assume all variation is due to differences in performance. With these confidence intervals taking uncertainty into account, and allowing for random variation, we can make much more robust statements about differential performance.

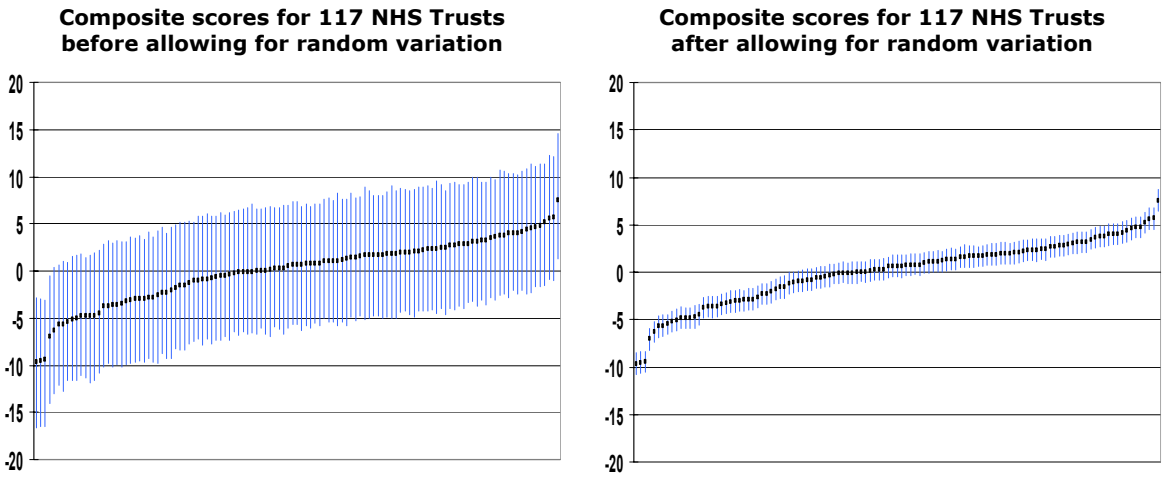
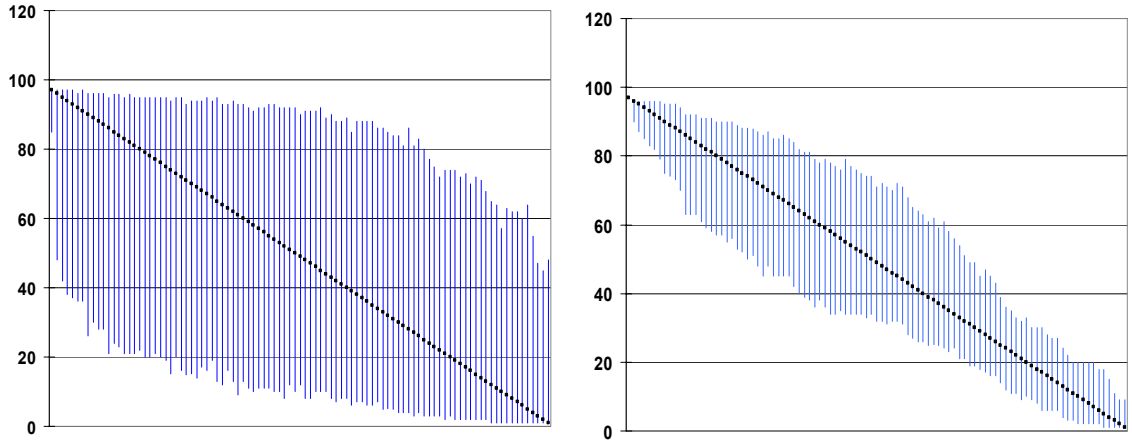


Figure 2: Composite score and 95% confidence interval for 117 NHS Trusts

Composite rankings for 97 Local Authorities before allowing for random variation

Composite rankings for 97 Local Authorities after allowing for random variation



Local Authorities ranked from worst to best (with uncertainty intervals)

Figure 3: Composite ranking and 95% confidence interval for 97 local authorities

When examining the rankings rather than scores of organisations, we obtain the results in Figure 3 for local authorities. Authorities are ranked from worst (97th) to best (1st) and the vertical lines show the 95% confidence intervals around these rankings. Again, the left panel illustrates the **high degree of uncertainty in the rankings** of authorities **prior to taking account of random variation** in the underlying performance indicators with almost all confidence intervals overlapping. The right panel illustrates that this naive view, attributing all variation to randomness, is altered radically (with still some overlap in the middle) after accounting for random variation. Results are similar when the analysis is repeated for hospitals.

2. Examining 'real' performance variation

A similar pattern is found in the degree of random variation in the underlying individual indicators across the two sectors. For healthcare it varies from 80% (inpatient waiting times) to 1% (sickness absence rates). For local authorities it ranges from 98% (unauthorised absences secondary schools – education) to 1% (older people helped to live at home – social services). It is evident that there is a similarly **wide range in the estimated proportion of random variation** across the 10 hospital and 35 local authority performance indicators. Managers will therefore have varying degrees of leverage to control certain types of performance indicators.

Some of the random variation may be driven by:

- subtle changes in the definitions of indicators over time,
- subtle changes in the way data is collected or measured over time,
- performance targets attached to individual indicators which may lead to increased variation within organisa-tions over time as they improve their performance, and
- possible "gaming" behaviour.

3. Examining alternative aggregation methods

We found that small **changes in methods to aggregate** underlying indicators to construct the composite indicator **can have a substantial impact**.

In the local government sector (under CPA), a differential weighting is applied to the various domains as set out in Table 2.

Table 2: Weighting applied to the Comprehensive Performance Assessment (CPA) data

Seven domains:	Weight:
Education (Ed)	4
Social services (SS)	4
Environment (Env)	2
Housing (Ho)	2
Libraries and leisure	1
Benefits	1
Use of resources	1

We explore the impact on the original composite indicator of changing the weights on the underlying performance indicators. Table 3 shows the impact of increasing and decreasing the weights on the performance indicators in education (Ed) and social services (SS) by a factor of 4 and increasing and decreasing the weights on the performance indicators in environment (Env) and housing (Ho) by a factor of 2. The final

column shows the impact of simultaneously amending the weights for the seven domains according to Table 2. The results highlight the change in ranking across 97 places for the top, middle and bottom 3 local authorities. The correlation between the new and original rankings varies between 0.81 and 0.96. The **largest jump in position** for an individual authority is 54 places, **more than half the league table**. On average, authorities change between 6 and 13 places in the rankings, depending on the changes made to the weighting system. Clearly, changes to the weighting structure of performance indicators can have a profound impact on the rankings of organisations. Results are again very similar when the analysis is repeated for hospitals.

We also illustrate that **using decision rules** to assign hospitals or local authorities to ordinal categories (e.g. 0-3 stars or 'excellent', 'good', 'poor') which are typically applied in the construction of composite scores, **produces even more variability in ratings**. Box 1 gives an example of a hypothetical set of decision rules which can be applied to hospitals to create a composite star rating.

Table 3: Rankings for local authorities after changing weights on underlying performance indicators

	Original	(Ed+SS) ×4	(Ed+SS) ×¼	(Env+Ho) ×2	(Env+Ho) ×½	(Ed+SS) ×4 + (Env+Ho) ×2 + Rest ×1
Top 3	1	3	4	1	9	9
	2	21	1	2	5	10
	3	2	13	11	1	4
Middle 3	48	58	43	51	49	75
	49	55	52	30	64	60
	50	54	54	62	47	55
Bottom 5	95	43	96	94	93	65
	96	97	76	84	96	96
	97	95	97	97	97	97
Correlations		0.81	0.88	0.91	0.96	0.88
Largest change		52	54	42	23	38
Average change		13	10	9	6	11

Box 1: Sequential decision rules applied to create a composite index

The indicators are first transformed into categorical variables (as for Star ratings) on a scale of 1 to 3.

Then:
3 star if achieve a score of 3 on certain variables
2 star if achieve a score a 2 or 3 on certain other variables
1 star if achieve a score of 1 or 2 on certain variables
0 star if achieve a score of 1 on certain other variables

Table 4: Proportion of times a sample of 10 hospitals receive a particular rating on the composite constructed from decision rules

Hospital	Original composite score	Percentage of times a hospital is given a score of:			
		0	1	2	3
A	0	**100**	0	0	0
B	0	**82**	18	0	0
C	0	**66**	0	34	0
D	1	2	**61**	0	38
E	1	0	**100**	0	0
F	2	38	2	**32**	28
G	2	19	0	**81**	0
H	2	0	0	**100**	0
I	3	38	2	0	**61**
J	3	0	0	44	**56**

Table 5: Proportion of times hospitals receive a particular rating on the composite constructed from decision rules

Number of hospitals	Original composite score	Percentage of times a hospital is given a score of:			
		0	1	2	3
33	0	**79.0**	8.6	11.9	0.4
22	1	7.7	**77.4**	8.2	6.7
46	2	9.8	10.0	**75.5**	4.7
16	3	9.1	17.7	19.5	**53.7**

Results in Table 4 show the high degree of uncertainty which is introduced by the use of decision rules. The table illustrates the frequency distribution for the number of times a hospital is placed in each of the four categories, using a sample of 10 hospitals. We find that whilst there is relative stability for the worst performing hospitals (with zero score), this is not the case for the best performers (three stars). So for example, hospitals *A* and *B* achieve a zero score in 100% and 82% of the simulations respectively; whereas hospitals *I* and *J* receive the score of 3 stars only 61% and 56% of the time respectively. Indeed, hospital *I* suffers a catastrophic relegation to a zero score in 38% of the simulations.

Table 5 summarises the results for all 117 hospitals and reiterates the **greater stability in the ranking of the worst hospitals** over the 1000 simulations. Results were similar for local authorities.

4. Examining composite indicators over time

Changes in ratings from year to year may arise due to genuine changes in performance reflected in fluctuations in the performance data, but also to changes in the methodology used to create them. Figure 4 explores both scenarios.

In the actual NHS Star ratings, the methodology (aggregation methods, weighting systems, and decision rules) can change every year. The first graph shows the changes in the actual NHS Star ratings over time for our sample of 117 hospitals. The figure illustrates the proportion of hospitals in each year which move up or down 1, 2 or 3 stars, or stay the same (zero change).

The second graph shows the changes in our generic composite rating over time for our sample of 117 hospitals. In our generic measure, the methodology remains unchanged over time. The proportion of hospitals whose ratings remain unchanged over time (zero) is generally much higher in our generic composite indicator – the smaller proportion of changes in ratings is purely due to changes in data relating to performance. Thus, if **methods to construct composites are constant over time**, there will be **much greater stability in ratings** over time.

Year on year changes to methods are often driven by changes in policy priorities and political imperatives and it is unrealistic to expect performance measurement systems to remain constant. Organisations may therefore need to learn to adapt to continual changes in the performance measurement system on which they are being rated. However, continual changes to the system could create difficulties in adopting longer term planning horizons and more strategic approaches to performance improvement.

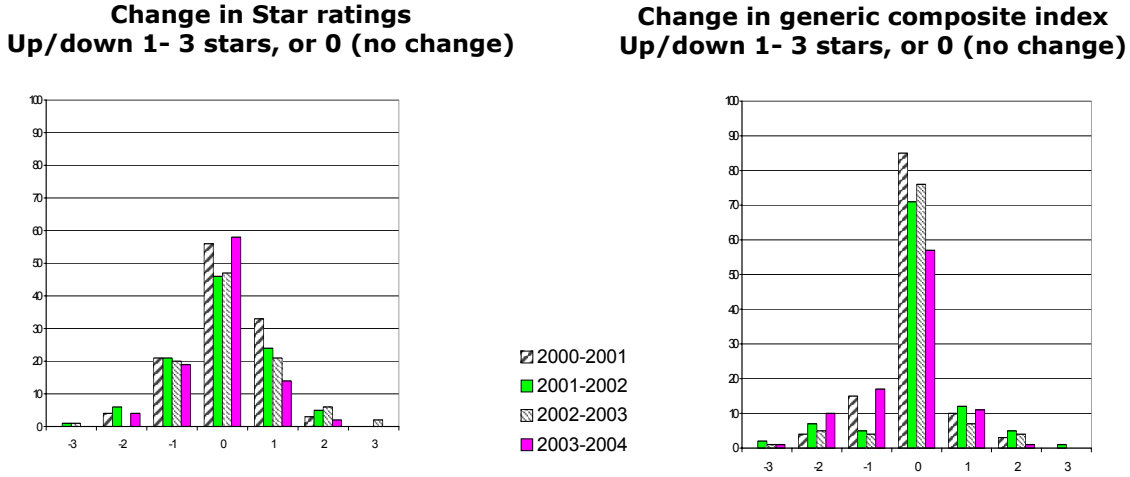


Figure 4: Change in Star ratings and generic composite index over time

What can we conclude? Implications for policy and practice

Composite performance measures are often attached to regulatory mechanisms whereby organisations are rewarded or punished according to the outcome of the composite indicator. The use and publication of composite performance measures can generate both positive and negative behavioural responses. If significant policy and practice decisions rest on the outcome of the composite, it is important to have a clear understanding of the potential risks involved in constructing a composite and arriving at a ranking. Key implications for policy and practice are:

1. Every performance measure consists of real variations in performance, and random fluctuations (natural statistical variation). In any performance benchmarking system, we need to know an estimate of the degree of random variation for each indicator so that we can draw definitive conclusions about real differences in performance. We **disentangle genuine performance variations** (for which managers can be held accountable) **from random fluctuation** in the measurement of performance indicators. Stripping out the variation for which managers can be held accountable, means we gain much greater precision in performance assessment, as uncertainty shrinks and we can produce much more robust composite performance measures.
2. The **construction of composite indicators is sensitive to methodological choices**. Changes in aggregation methods (either altering weightings or decision rules) can have a substantial impact on results, with organisations jumping from one end of the league table to the other following small alterations in the aggregation rules.
3. The **choice of a weighting system can have a significant impact** on the rankings of individual units within the composite. The choice of weights may be ad hoc and arbitrary with a lack of consideration for whose preferences the weights reflect and how robust these are. Greater attention should be paid to the origin and nature of weights and the sensitivity of composites to changes in the weighting structure.
4. **"Decision rules" need to be treated with caution**. Subtle and highly subjective changes to the decision rules can impact dramatically on the composite index and rankings of organisations.
5. In addition to random variation causing uncertainty in performance assessment, **year-to-year changes to the methodology can have a major impact** on the stability of performance ratings. Thus organisations may jump around the league table because of annual changes in political priorities which may present a misleading picture of performance over time.
6. The proper treatment of uncertainty in composite performance measures is crucial - **composites need to be published with indications of uncertainty** to communicate the sensitivity of the reported measure.

References

1. Freudenberg, M. (2003) *Composite indicators of country performance: A critical assessment*, OECD STI Working paper DSTI/DOC 2003/16, OECD: Paris.

2. Smith, P. (2002) Developing composite indicators for assessing health system efficiency, in Smith, P.C. (ed.) *Measuring up: Improving the performance of health systems in OECD countries*, OECD: Paris.

3. Joint Research Centre (2002) *State of the art report on current methodologies and practices for composite indicator development*, Report prepared by the Applied Statistics Group, Institute for the Protection and Security of the Citizen: European Commission, June 2002.

4. Mooney, C.Z. (1997) *Monte Carlo simulation*, A Sage University Paper, No. 116 in Series: Quantitative Applications in the Social Sciences, Sage Publications Ltd: London.

This research was funded by the Economic and Social Research Council (ESRC) grant number RES-153-25-0031 under the Public Services Programme. Full details of the Public Services Programme are available here: http://public-services.politics.ox.ac.uk/

Full details of the research are available here: Jacobs, R., Godddard, M. and Smith, P.C. (2006) *Public Services: Are composite measures a robust reflection of performance in the public sector*, Centre for Health Economics Research Paper 16, University of York. http://www.york.ac.uk/inst/che/pdf/rp16.pdf