THE UNIVERSITY of York

CHE
Centre For Health Economics

Public Services: Are Composite Measures a Robust Reflection of Performance in the Public Sector

CHE Research Paper 16

PLEASE DO NOT QUOTE WITHOUT THE AUTHORS' PERMISSION

Public Services:

Are composite measures a robust reflection of performance in the public sector?

Rowena Jacobs, Maria Goddard, Peter C. Smith,

Centre for Health Economics
University of York
York, Y010 5DD
United Kingdom
Tel: +44-1904-321425
Fax: +44-1904-321454
Email: rj3@york.ac.uk

June 2006

Background

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The new CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Disclaimer

Papers published in the CHE Research Paper (RP) series are intended as a contribution to current research. Work and ideas reported in RPs may not always represent the final position and as such may sometimes need to be treated as work in progress. The material and views expressed in RPs are solely those of the authors and should not be interpreted as representing the collective views of CHE research staff or their research funders.

Further copies

Copies of this paper are freely available to download from the CHE website www.york.ac.uk/inst/che/pubs. Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321458 for further details.

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/inst/che

Abstract

A composite indicator is an aggregated index comprising individual performance indicators. Composite indicators integrate a large amount of information in a format that is easily understood and are therefore a valuable tool for conveying a summary assessment of performance in priority areas.

This research investigates the degree to which composite measures are an appropriate metric for evaluating performance in the public sector. Do they reflect accurately the performance of organisations? To what degree are they influenced by the uncertainty surrounding underlying indicators on which they are based? Are they robust and stable over time? The construction of composite measures creates specific methodological challenges that make such questions especially pertinent. We address these through a series of quantitative analyses of panel data relating to healthcare (Star ratings of NHS acute Trusts) and local government (Comprehensive Performance Assessment (CPA) ratings of authorities) in England where composites have been widely used.

The creation of a composite comprises a number of important steps, each of which requires careful judgement. These include the specification of the choice of indicators, the transformation of measured performance on individual indicators, the specification of a set of weights on individual indicators, and combining the indicators using aggregation methods or decision rules. We use Monte Carlo simulations to examine the robustness of performance judgements to these different technical choices. We show the extent to which composites provide stable performance rankings of organisations over time and assess whether variations are due to genuine performance improvement or merely the result of random statistical variation.

The analysis suggests that the judgements that have to be made in the construction of the composite can have a significant impact on the resulting score. Technical and analytical issues in the design of composite indicators have important policy implications. We highlight the issues which need to be considered in the construction of robust composite indicators so that they can be designed in ways which will minimise the potential for producing misleading performance information which may fail to deliver the expected improvements or even induce unwanted side-effects.

Keywords: performance measurement, performance indicators, composite indicators

Table of contents

1. INTRODUCTION 1
2. ARGUMENTS FOR AND AGAINST THE USE OF COMPOSITE INDICATORS 2
3. THE DEVELOPMENT AND USE OF COMPOSITE INDICATORS IN THE PUBLIC SECTOR 3
3.1. THE STAR RATING SYSTEM FOR NHS PROVIDER TRUSTS 3
3.1.1. Key targets 3
3.1.2. 'Balanced scorecard' indicators 4
3.1.3. Clinical Governance Review 5
3.1.4. Determining Star ratings 6
3.2. THE COMPREHENSIVE PERFORMANCE ASSESSMENT FOR LOCAL GOVERNMENT 7
3.2.1. Housing, Environment, Use of Resources and Libraries and Leisure 10
3.2.2. Benefits 13
3.2.3. Social services 14
3.2.4. Education 15
4. EMPIRICAL ANALYSIS 17
4.1. THE DATA 17
4.2. METHODS AND RESULTS 17
4.2.1. Choosing the indicators 17
4.2.2. The chosen indicators 18
4.2.3. Examining uncertainty 24
4.2.4. Random variation on performance indicators 29
4.2.5. Introducing changes in weights 35
4.2.6. Introducing decision rules 38
4.2.7. Changes in ratings over time 42
5. CONCLUSIONS 46
6. REFERENCES 48

List of tables

Table 1: Rules for summary key target score by total penalty points 4
Table 2: Confidence interval method threshold for scoring clinical performance indicators 5
Table 3: Percentile method thresholds for scoring non-clinical performance indicators 5
Table 4: Rules for allocating star rating - acute and specialist trusts 6
Table 5: Matrix of performance rating based on performance against key targets..7
Table 6: The number of Best Value Performance Indicators (BVPIs) used in the Comprehensive Performance Assessment (CPA) 8
Table 7: Determination of scorecard for Comprehensive Performance Assessment (CPA)...9
Table 8: The percentile approach for scoring performance indicators 10
Table 9: Averaging the percentiles for the service block: Use of Resources - Financial Administration...10
Table 10: Converting average percentiles to scores.....................................11
Table 11: Adjusted thresholds when there are fewer than 4 PIs....................11
Table 12: Thresholds for determining the service assessment score for environment, housing and culture...12
Table 13: Approach to scoring performance indicator data for housing, environment and culture 12
Table 14: Benefit Fraud Inspectorate assessments within the CPA...............13
Table 15: Star ratings for social services performance 15
Table 16: Operation of performance and improvement rules........................16
Table 17: Example of Local Education Authority CPA education scorecard (for the City of York) 16
Table 18: Variable names and descriptions, healthcare 19
Table 19: Descriptive statistics of variables used to form composite, healthcare 19
Table 20: Variable names and descriptions, local government......................20
Table 21: Descriptive statistics of variables used to form composite, local government...21
Table 22: Descriptive statistics of standardised variables and new composite, healthcare 23
Table 23: Descriptive statistics of standardised variables and new composite, local government...23
Table 24: Descriptive statistics of standardised variables from simulations (1000 replications), healthcare...25
Table 25: Descriptive statistics of standardised variables from simulations (1000 replications), local government 25
Table 26: Descriptive statistics of standardised variables and new composite from simulations (1000 replications), healthcare...26
Table 27: Descriptive statistics of standardised variables and new composite from simulations (1000 replications), local government 26
Table 28: Proportion of random variation on underlying performance indicators, healthcare...30

Table 29: Proportion of random variation on underlying performance indicators, local government 31
Table 30: Impact of different weighting systems on hospitals' composite rankings (with 95% uncertainty intervals) for top, middle and bottom 5 ranked hospitals 36
Table 31: Weighting applied in the Comprehensive Performance Assessment (CPA) 37
Table 32: Impact of different weighting systems on local authorities' composite rankings for top, middle and bottom 5 ranked authorities 37
Table 33: Sequential decision rules applied to create the illustrative composite measure, healthcare 38
Table 34: Sequential decision rules applied to create the illustrative composite measure, local government 39
Table 35: Frequency distribution of a sample of 10 hospitals on the new composite index constructed from decision rules 41
Table 36: Frequency distribution of a sample of 15 local authorities on the new composite index constructed from decision rules 41
Table 37: Frequency distribution of hospitals on the new composite index constructed from decision rules 42
Table 38: Frequency distribution of local authorities on the new composite index constructed from decision rules 42
Table 39: Data availability over time for the 10 chosen indicators, healthcare 43
Table 40: Data availability over time for the 35 chosen indicators, local government 43

List of figures

Figure 1: Performance of PIs against relevant thresholds 12
Figure 2: Uncertainty intervals around composite score using simulations, healthcare 27
Figure 3: Uncertainty intervals around composite score using simulations, local government...28
Figure 4: Uncertainty intervals around composite rank using simulations, healthcare 28
Figure 5: Uncertainty intervals around composite rank using simulations, local government 29
Figure 6: Example of variance around survival rates with 1000 observations with distribution N(0,1) and N(0,0.27) 32
Figure 7: Uncertainty intervals around composite score using simulations with random variation taken into account, healthcare 33
Figure 8: Uncertainty intervals around composite score using simulations with random variation taken into account, local government 33
Figure 9: Uncertainty intervals around composite rank using simulations with random variation taken into account, healthcare 34
Figure 10: Uncertainty intervals around composite rank using simulations with random variation taken into account, local government 34
Figure 11: Uncertainty intervals around a new composite score constructed from thresholds, healthcare 40
Figure 12: Uncertainty intervals around a new composite score constructed from thresholds, local government 40
Figure 13: Frequency distribution of hospitals changing ratings on the actual star ratings, n=117 44
Figure 14: Frequency distribution of hospitals changing ratings on the new composite index constructed from decision rules, n=117.............................45
Figure 15: Frequency distribution of local authorities changing ratings on the actual CPA ratings, n=97 45
Figure 16: Frequency distribution of local authorities changing ratings on the new composite index constructed from decision rules, n=97.........................46

Executive summary

1. This research investigates the degree to which composite measures are an appropriate metric for evaluating performance in the public sector. Composite indicators are an aggregation of a number of underlying performance indicators. They provide a single score or rating that is easy to understand. Their use is widespread as they offer an attractive way of summarising a wealth of performance data. But do they reflect accurately the performance of the organisation?

2. The construction of composite measures is not straightforward and the associated methodological challenges can create the potential for composite measures to be misinterpreted or manipulated. Moreover, the dangers are not restricted to technical issues as the use and publication of composite performance measures can generate both positive and negative behavioural responses, so careful consideration needs to be given to their creation and subsequent use.

3. The aims of this research were to highlight the methodological challenges underpinning the construction of composite measures of performance in the public sector.
   - To what degree are composite indicators influenced by the uncertainty surrounding underlying performance indicators?
   - To what degree does random variation in the measurement of the underlying performance indicators impact on the composite?
   - How robust are the composite performance scores and rankings of organisations?

4. We use longitudinal data from the healthcare and local government sectors in England to examine these issues. The purpose was not to replicate the construction of the two types of composite ratings that currently exist in these sectors (namely the NHS Star Ratings for hospital Trusts and the Comprehensive Performance Assessment (CPA) for local authorities), but rather to use the two datasets to explore the research questions posed above.

5. Our research methodology was:
   - To construct our own generic composite indicator for each sector and use these to examine the sensitivity of rankings of organisations to various methodological issues. We used Monte Carlo simulation methods, a stochastic modelling approach, to simulate the construction of the composite indices. This enabled us to calculate the degree of uncertainty around the composite by examining the range of performance scores a single organisation obtains over 1000 simulations.
   - To explore the degree to which the indicators are subject to random variation (measurement error and natural variation). This enabled us to disentangle the random variation from “true” differences in performance.
   - To compare the composite indices in the two sectors.
   - To explore changes in the two composite indicators over time and assess how much stability there is in ratings over time.

6. Our results produced a set of rankings of hospitals and local authorities according to their composite score with a 95% confidence interval around this composite score arising from the simulations. These confidence intervals reflected a considerable degree of uncertainty in both the composite score and ranking of organisations as the confidence intervals overlap over the entire range of the composite. This casts doubt on whether one organisation is performing better than another and leads to concerns over the ranking of organisations on the basis of the composite.

7. We decompose the variation on the performance indicators (to separate out real differences in performance from those due to random variation, measurement error, sampling error or simply natural variation. After taking account of random variation in the underlying performance indicators, we were able to estimate genuine performance variations with much greater precision. It is then possible to say with certainty that (for

example) the hospitals in the bottom quartile of a league table are scoring less well than those in the top quartile.

8. We illustrate that small changes in methods to construct composite indicators can have a substantial impact on the results. For example, when varying weights attached to the underlying indicators for aggregation, hospitals and local authorities jump around considerably in their rankings in the league table. 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. For example, a hospital might be a 3 star for 50% of the simulations, but then receive a 0 star for nearly 40% of the simulations.

9. Since the popularity of composite indicators is likely to endure, we conclude that it is essential they be published with reasonable indications of uncertainty, in order to communicate the sensitivity of the measure.

10. The key contributions of this work are the disentangling of genuine performance variations from random fluctuation in the measurement of performance indicators and the clear visual illustrations of the impact of uncertainty on performance measures. These findings provide a mechanism whereby researchers and analysts can account for 'random variation' in performance indicators more generally.

11. This research is particularly relevant to those concerned with assessing performance and designing systems in public services. This research indicates that methodological issues in the design of composite indicators are of interest not merely from a technical perspective, but because they also have important policy implications. The findings will have a very real impact in practice if performance is assessed using composite indicators and rewards or penalties are given on the basis of these results.

1. Introduction

In this research we investigate the degree to which composite indicators are an appropriate metric for evaluating performance in the public sector. Composite indicators are an aggregation of a number of underlying performance indicators. They provide a single score or rating that is easy to understand and offer an attractive way of summarising a wealth of performance data in priority areas. They are increasingly being used to measure the performance of, and also to rank, organisations and institutions in economic, social and policy areas, in both the public and private sector (Freudenberg, 2003). International interest in the use of composite measures is widespread – for example, the OECD and the European Commission Joint Research Centre have investigated aspects of the methodology and use of composite indicators in many different areas (Joint Research Centre, 2002).

One might question why there is such a compelling need to produce a summary indication of performance in the form of a composite score. Possible arguments in favour of creating composites include focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the 'big picture' in a way in which the public can understand. In contrast to piecemeal indicators based on individual performance measures, they can offer policymakers a summary of complex or multi-dimensional issues. They provide an attractive option for accountability purposes as it is relatively easy to track progress over time using a single indicator rather than a whole package of indicators. Although there is a range of counter-arguments (which we discuss in section 2), the temptation to summarise complex processes into a single figure for benchmarking organisational performance seems irresistible.

Current government policy in England emphasises the creation and publication of composite performance measures in the public sector and they are used widely in health, social services, education, universities, local government and other service areas. They are often used to create league tables and rankings of performance. Although policy developments may shift the scope and content of such measures, the search to transform a plethora of disaggregated performance data into simple scores or ratings is likely to be enduring.

Despite the proliferation of composite indicators across various sectors, the construction of a composite indicator is not straightforward methodologically. This leaves it open to misinterpretation and potential manipulation. Questions of the accuracy, reliability and appropriateness of such indices, need to be addressed if major policy, financial and social decisions hinge on the results of composite indicators.

Little is known about the degree to which composite measures are an appropriate metric for evaluating performance in the public sector. Do they reflect accurately the performance of the organisation? To what degree are they influenced by the uncertainty surrounding the underlying indicators on which they are based? Are they robust and stable over time? Many of these issues are of course common in considering any type of performance indicator. However, the use of composite measures compounds these difficulties and is associated with additional methodological challenges that influence the degree to which they may represent an adequate measure of performance. In particular, although apparently simple, the process of creating the composite from a wealth of disparate performance data is very complex and involves a series of judgements at each stage of construction.

The steps involved in constructing a composite include:

- choosing the entities to be assessed
- choosing the organisational objectives to be encompassed in the composite
- choosing the indicators to be included
- transforming measured performance on individual indicators
- combining the individual measures using addition or some other decision rules
- specifying an appropriate set of weights
- adjusting for environmental or other uncontrollable influences on performance
- adjusting for variations in expenditure if a measure of efficiency is required
- using sensitivity analysis to test the robustness of the composite to the various methodological choices

This report examines the robustness of performance judgements to different technical choices made at each step in the construction of a composite. We use longitudinal data from the healthcare and local government sectors in England. The purpose is not to replicate the construction of the two types of composite ratings that currently exist in these sectors (namely the NHS Star Ratings for hospital Trusts and the Comprehensive Performance Assessment (CPA) for local authorities), but merely to use the ratings data to explore the generic issues involved in constructing a composite index.

The objectives of this study are:

1. to investigate the degree to which composite measures provide a robust summary of performance of public sector organisations (specifically, health and local government),
2. to examine how much of the variation in performance indicators in the public sector is due to random variation (or measurement error and natural statistical variation), rather than genuine variations in performance, and how this impacts on the construction of composite indices,
3. to compare composite performance measures in the public sector,
4. to examine alternative methodologies for constructing composite measures of performance, and
5. to explore changes in composites over time and assess how much stability there is in performance ratings of organisations over time.

Section 2 of the report presents arguments for and against the use of composite performance indicators, while section 3 describes the use of composite indicators in health care (Star Ratings) and local government (CPA), highlighting the methodology employed in their construction. The empirical work follows in section 4 where we address each of the above research objectives, and we conclude in section 5 with some analytical and policy conclusions.

2. Arguments for and against the use of composite indicators

Composite indicators have a high profile in the media and potentially have an important role alongside the publication of individual performance indicators.

Some of the arguments for developing composite indicators include the following (Smith, 2002):

1. They place performance at the centre of the policy arena
2. They can offer a rounded assessment of performance
3. They enable judgements to be made on system efficiency
4. They facilitate communication with the public and promote accountability
5. They indicate which organisations represent beacons of best performance
6. They indicate priority organisations for improvement efforts
7. They may stimulate the search for better analytical methods and better quality data
8. They present the ‘big picture’ and can be easier to interpret than trying to find a trend in many separate indicators.

On the other hand, composite indicators may lead to a number of dysfunctional consequences and there are a number of arguments against their use including the following (Smith, 2002):

1. By aggregating individual performance measures, composites may disguise serious failings in some parts of the system
2. As measures become aggregated it becomes more difficult to determine the source of poor performance and where to focus remedial action
3. The individual performance measures used in the composite are often contentious
4. A composite that is comprehensive in coverage may have to rely on poor quality data in some dimensions
5. A composite that ignores some performance measures because they are difficult to measure may distort behaviour in undesirable ways
6. The composite measure will depend crucially on the weights attached to each performance dimension. However, the methodology by which weights are elicited and decisions on whose preferences they should reflect are unlikely to be straightforward.

3. The development and use of composite indicators in the public sector

3.1 The star rating system for NHS provider Trusts

The NHS Performance Assessment Framework (PAF), published in April 1999, introduced a new broader-based approach to assessing performance in the NHS by encouraging action across six areas (Health improvement; Fair access; Effective delivery of appropriate health care; Efficiency; Patient/carer experience; and Health outcomes of NHS care). The PAF was supported by the publication in June 1999 of the first set of High Level Performance Indicators (HLPIs) and Clinical Indicators (CIs) for both Health Authorities and NHS Trusts respectively (Department of Health, 2000). This was the first full range of indicators for NHS hospital Trusts and gave impetus to the process of publishing information on the performance of NHS organisations in order to provide comparisons and improve performance overall.

In September 2001, the first set of performance star ratings were published by the Department of Health for acute NHS Trusts 2000/01 (Department of Health, 2001). The Star ratings are a composite index score given to each NHS organisation which are supposed to provide an overall assessment of performance across a number of indicators. In July 2002, the second set of Star ratings were published by the Department of Health, now covering acute Trusts, specialist Trusts, ambulance Trusts and indicative ratings for mental health Trusts for 2001/02 (Department of Health, 2002a). Primary Care Trusts (PCTs) received a separate publication, describing their performance against a range of suitable indicators, but not a rating. In July 2003, the next set of Star ratings were published, covering again all types of NHS Trusts and PCTs. In this round, the Commission for Health Improvement (CHI), took over responsibility for performance ratings and indicators from the Department of Health (Commission for Health Improvement, 2003a). Two more sets of Star ratings were published in July 2004 and July 2005 respectively, by the new regulator of performance, the Healthcare Commission (Healthcare Commission, 2004; Healthcare Commission, 2005).

The NHS Performance Ratings system places NHS Trusts in England into one of four categories from the highest levels of performance (three stars) to the poorest levels of performance against the indicators (zero stars).

The methodology for the five years of Star ratings has remained relatively constant, although some important changes have been made to the individual indicators covered and how they are aggregated together. The Trust Star ratings comprise similar areas of performance to the PAF which, taken together, should give a balanced view of the performance of NHS hospital Trusts. There were broadly four areas of indicators in 2000/01: Clinical effectiveness and outcomes; Efficiency; Patient/carer experience; and Capacity & capability. In the last four years of Star ratings the key areas have been:

- Key government targets;
- The ‘balanced scorecard’ covering 3 focus area of Clinical focus; Patient focus; and Capacity & capability.

3.1.1 Key targets

Performance against key targets is assessed in terms of whether the target has been achieved, whether there has been some degree of underachievement or whether the target was significantly underachieved. Trust performance is considered to be of concern if there are:

- a number of targets with some degree of underachievement
- a smaller number of targets with significant levels of underachievement
- some combination of both

The scores are presented in terms of the categories: “achieved”, “under achieved” and “significantly under achieved” respectively.

For each key target, a Trust is allocated penalty points in relation to its performance level using the following rules:

- Achieved target: 0 points
- Underachieved: 2 points
- Significantly underachieved: 6 points

Thus, significantly underachieving on one key target is equivalent to underachieving on three key targets, reflecting the seriousness of failure to meet the target.

The overall results on key targets is summarised on a 4-point scale. The method of aggregation is based on a system of penalty points for each key target. These are added together to give a total penalty score on key targets, with each key target equally weighted. Then the rules shown in the following table are applied to obtain the final aggregate key target score for each Trust (Healthcare Commission, 2005).

Table 1: Rules for summary key target score by total penalty points

Key target score	Acute Trusts
3 Pass	<=2
2 Borderline	3 - 6
1 Moderate Fail	7 - 12
0 Fail	>=13

Source: Healthcare Commission (2005) http://ratings2005.healthcarecommission.org.uk/Downloads/methodology_acute.doc

In determining whether a Trust has met thresholds on key targets, a Trust's value on any one indicator is rounded to a defined level of precision for that indicator. The precision of the rounding is based on the scale of indicator values. This approach was adopted to allow some margin of error for a Trust with an indicator value close to the threshold. For example, for an indicator with values ranging between 30% and 100% and a threshold set at 80%, a 2 decimal points rounding would be given; while for an indicator with values between 0 to 5% and a threshold set at 1.5%, a 3 decimal points rounding is applied.

A certain number of indicators are measured as a percentage value. This can give rise to a problem for Trusts with small denominators (low service volumes), especially when the value scale and thresholds tend to be very small. This issue affects, for example, the inpatient and outpatient waiting time breach key targets. To counter this problem, a "low activity rule" was adopted for these two indicators in the 2004 rating. A similar low activity rule was adopted for the key target "All cancers: 2 week wait". If a Trust had seen a small number of patients (less than 10 in the denominator), then 'data not available' is applied.

3.1.2 'Balanced scorecard' indicators

The results on 'balanced scorecard' indicators are first summarised into an overall score for each focus area. A 3-band scoring system (0,1,2) is used for each focus area.

All indicators are equally weighted within their focus area by simply summing up indicator band scores to produce the total score for the Trust. These aggregated scores are then split into 3 bands (0,1,2) according to specific thresholds for each focus area.

Overall scores are awarded as follows:

- in the top band (equal or above 45th percentile): two points
- in the middle band (17th-45th percentile): one point
- in the bottom band (below 17th. percentile): zero points

These overall scores for each focus area are added together to give a score for the 'balanced scorecard' on a scale of (0 to 6) where a score of 6 indicates the highest level of performance. All indicators are equally weighted within their focus area in such a way as to ensure that despite differing numbers of indicators, each focus area carries the same weight.

A Trust's performance against a 'balanced scorecard' indicator is measured using a banding system, with a score of between 5 (good) and 1 (poor). Underlying performance indicators are transformed from continuous variables into categorical variables of either 3 or 5 categories. Various banding methods are applied to convert indicator values into one of the five bands: the statistical confidence interval method; the absolute thresholds method; and the mapping score method.

The clinical indicators are published with confidence intervals which means that performance is split into three bands dependent on whether the organisation's confidence interval overlaps the England average for the indicator, as shown in the following table.

Table 2: Confidence interval method threshold for scoring clinical performance indicators

Overlap with England confidence interval:	**Band given:**	**Labelled as:**
No overlap and organisation is worse than England average confidence interval	1	Significantly below average
Overlap, showing organisation is similar to England average confidence interval	3	Average
No overlap and organisation is better than England average confidence interval	5	Significantly above average

Source: Department of Health (2002b) http://www.doh.gov.uk/performanceratings/2002/method_acute.html

In earlier years, the thresholds for deciding the cut-offs for banding performance indicators were not necessarily the same for each variable. The default position was that in the following table which simply splits performance into 5 bands by percentile.

Table 3: Percentile method thresholds for scoring non-clinical performance indicators

Performance falls in:	**Band given:**	**Labelled as:**
1st to 10th percentile	1	Significantly below average
11th to 30th percentile	2	Below average
31st to 70th percentile	3	Average
71st to 90th percentile	4	Above average
91st to 100th percentile	5	Significantly above average

Source: Department of Health (2002b) http://www.doh.gov.uk/performanceratings/2002/method_acute.html

The absolute threshold method of banding was introduced in 2004 and replaces the relative percentile method shown above. Absolute thresholds are now set for either 3-bands (1,3,5) or 5-bands (1 to 5). This effectively removes the relative measurement on indicator banding applied in previous years so that a Trust's banding score on an indicator will not be decided by the performance levels of all other Trusts anymore.

An analogous approach to the key targets is adopted for determining whether a Trust has met a threshold. A Trust's indicator value is rounded to the defined level of precision for the threshold.

For indicators which only take a few number of ordinal values (e.g. 0, 1, 2), a mapping score method is used to define directly a mapping between the indicator values and the performance bands. This was introduced in the 2003 ratings for traffic light indicators (green, amber, red) and yes/no indicators.

3.1.3 Clinical Governance Review

A Clinical Governance Review (CGR) has been used in previous years to review performance across seven components:

1. risk management
2. clinical audit
3. research and education
4. patient involvement
5. information management
6. staff involvement
7. education, training and development

Each component is scored from I to IV. After each review, the Trust prepares an action plan to address areas for improvement identified by the CGR report. Any significant improvements are taken into account in calculating the Star rating.

The role of the CGR has evolved over time. In 2001, the assessment for Star ratings required only that the organisation had not received a critical review. However, at that time, just 16 acute Trusts and two mental health Trusts had undergone a CHI review. By contrast, when the 2003 Star ratings appeared, CHI reviews had been published for 160 acute Trusts, 28 mental health Trusts, 27 ambulance Trusts and 28 primary care Trusts. The 2003 NHS Star ratings for acute, specialist and mental health trusts were adjusted in accordance with the ‘Finsbury rules’ (Commission for Health Improvement, 2003b). In essence, these involved zero-rating any organisation that is evaluated as having achieved only the lowest standard of performance (level ‘I’) in five or more out of the seven areas of clinical governance assessed, apparently irrespective of the organisation’s performance on key targets or the scorecard. Three stars are awarded only to organisations that had achieved key targets, a balanced scorecard, at least three ‘III’s and no ‘I’s in the CHI review (Commission for Health Improvement, 2003b).

In 2004, CGRs were not incorporated into Star ratings for acute and specialist Trusts since the majority of reviews were by then several years old. They did however apply to other types of Trusts where the review was completed before March 2004 (Healthcare Commission, 2004). In February 2005, the Healthcare Commission made the decision to exclude all CGRs (or follow-up plans) from the performance ratings of all organisations in 2004/05 (Healthcare Commission, 2005).

3.1.4 Determining Star ratings

A complex process is used whereby a sequential set of decisions on achievement on the various key targets and ‘balanced scorecard’ indicators determines the ultimate star rating outcome. The most important driving factors for obtaining the highest rating are the key targets and (in earlier years) the CGR which entered both first and again at the very end of the set of decision steps for Trusts to achieve. It thus implicitly was given the highest weighting in the determination of the Star rating.

In the most recent ratings the following decision matrix is used. The general principles of this star allocation table are that the key targets play a dominant role and a Trust must perform well on both key targets and the ‘balanced scorecard’ to get 3 stars.

Table 4: Rules for allocating star rating - acute and specialist trusts

	Balanced Scorecard						
Key targets	BS=0	BS=1	BS=2	BS=3	BS=4	BS=5	BS=6
Fail: penalty points greater than 12	0 star	0 star	0 star	0 star	0 star	0 star	0 star
Moderate Fail: penalty points between 7 and 12 inclusive	0 star	1 star	1 star	1 star	1 star	1 star	1 star
Borderline: between 3 and 6 points inclusive	1 star	1 star	1 star	1 star	2 star	2 star	2 star
Pass: no more than 2 points	1 star	2 star	2 star	2 star	2 star	3 star	3 star

Source: Healthcare Commission (2005) http://ratings2005.healthcarecommission.org.uk/Downloads/methodology_acute.doc

In earlier years, when CGR also played a key role, the following decision algorithm was used:

Step 1
Does the CGR show significant weaknesses (calculated by achieving five or more scores of I across the seven components of a CGR)?
If YES – Rating is zero stars
If NO – Continue to Step 2

Step 2
The rating is then calculated using performance against the key targets. The number of key targets achieved and significantly underachieved is used to determine the level of rating possible. This is detailed in the table below:

Table 5: Matrix of performance rating based on performance against key targets

Key targets significantly underachieved	**3**				Zero stars	Zero stars	Zero stars
	2			One star	Zero stars	Zero stars	Zero stars
	1		Go to step 3	One star	One star	One star	Zero stars
	0	Go to step 4	Go to step 4	Go to step 3	Go to step 3	One star	One star
		9	**8**	**7**	**6**	**5**	**4**
		Key targets achieved					

Source: Department of Health (2002b)

Step 3
This step assesses the Trust's performance on the balanced scorecard for Trusts with moderate underachievement on the key targets. Is the Trust's performance either outside the lowest 20% for all three areas and within the top 50% for one area, or in the top 50% for two of the three areas?
If YES – Rating is two stars
If NO – Rating is one star

Step 4
This step assesses the Trust's performance on the balanced scorecard, for Trusts with little or no underachievement on the key targets. Does the CGR show some strength and no weaknesses (calculated as one score of III, and no scores of I across the seven components)?
If No – Rating is two stars
If YES – Continue to Step 5

Step 5
Is the Trust's performance on the balanced score card outside the lowest 20% in all three areas and in the top 50% in two out of three areas?
If YES – Rating is three stars
If NO – Continue to Step 6

Step 6
Is there a current CGR showing significant strengths and no weaknesses?
If YES – Rating is three stars
If NO – Rating is two stars

There are many decisions in the process which may impact on how many Trusts are accorded a particular rating in each category, in particular, the choice of performance indicators, the decisions on the thresholds for transforming the performance indicators into categorical variables, the decision rules which are applied in the matrix and the resultant implicit weighting given to each of the (groups of) indicators. The sensitivity to these choices is explored in this report.

The Star rating outcome has a significant reward and penalty schedule attached to it since Trusts which obtain a three star rating for a consecutive number of years may apply for Foundation Status which gives them significant financial and managerial freedoms and autonomy from central involvement. The incentives for ensuring a good outcome on the Star rating are therefore very strong.

The ratings are intended to be 'not primarily a commentary on the quality of clinical care', but rather to assess the 'overall patient experience' (Department of Health, 2001). However, both the construction and the impact of Star ratings have been questioned (Kmietowicz, 2003; Cutler, 2002; Snelling, 2003; Miller, 2002). Many of the concerns that have been raised on their construction are considered in this study.

3.2 The Comprehensive Performance Assessment for local government

Besides the health sector, a number of other public sectors in the UK use composite indicators in their performance rating systems. For example, the Audit Commission undertakes a Comprehensive Performance Assessment (CPA) of local government that covers a number of core services and draws on approximately 100 performance indicators (DTLR, 2001). CPA has been published in December of each year 2002, 2003, 2004 and 2005. This is an extremely complex composite indicator, since it is built up from a number of underlying composite indices measuring the key services.

The key services which are assessed are the following seven domains:

1. Benefits
2. Education
3. Environment
    a. *Transport*
    b. *Waste management*
    c. *Planning*
4. Housing
    a. *Housing management*
    b. *Community Housing*
5. Libraries & Leisure
6. Social services
    a. *Children's Services*
    b. *Adults' Services*
7. Use of resources

The CPA is performed by the Audit Commission who also cover the performance assessment for the domains environment, housing, libraries and leisure and use of resources (Audit Commission, 2003a; Audit Commission, 2003b). Star ratings for education, social services and benefits are produced as part of the CPA process and contribute to the overall CPA assessment. Education assessments are made by the Office for Standards in Education (Ofsted). Social care assessments are made by the Social Services Inspectorate (SSI) and benefits assessments are made by the Benefits Fraud Inspectorate (BFI).

This section describes how the seven domains (assessed by the four different inspectorates) are combined into the overall CPA scorecard for each council. Subsequent sections describe the construction of the underlying composites for the seven key areas. These all have very large similarities to the Star ratings in the NHS.

The CPA is an assessment of each authority’s performance across the full range of its services, its financial integrity, capacity to drive improvement, its underlying systems and processes and the way in which it relates to its community and partners. The scorecard summarises the performance of every assessed council. The overall CPA judgement draws information from a range of Government inspectors reports, Best Value Performance Indicators (BVPIs), audits, and assessments of service plans.

Table 6: The number of Best Value Performance Indicators (BVPIs) used in the Comprehensive Performance Assessment (CPA)

Corporate Health	16
Education	18
Social Services	14
Housing	9
Housing Benefit and Council Tax	6
Waste	8
Transport	9
Planning	5
Environmental Health and Trading Standards	1
Culture Services / Libraries and Museums	4
Community Safety	7
Community Legal Service	1
Fire	11

Source: (ODPM, 2003)

Key services then receive a score ranging from 1 (the lowest) to 4 (the highest). The Audit Commission also assesses how the council is run and how the council is going about delivering its priorities, again using a score of 1 to 4. These two scores are combined to reach the overall score of excellent, good, fair, weak or poor, which is known as 'the scorecard'.

Table 6 shows the approximate number of Best Value Performance Indicators used in the CPA.

Because there are differences in the size of councils and the types of services they deliver, two streams of assessment have emerged, namely Single Tier and County Councils and District Councils. Although these differ in the details, the overall approaches are similar (Audit Commission, 2003c).

The scores from all the services are combined according to the following process (for Single Tier and County Councils respectively) as set out in the following table.

Table 7: Determination of scorecard for Comprehensive Performance Assessment (CPA)

1. Current scores on services range from 1 (poor) to 4 (excellent)		
2. Weighting between services on the seven domains is:		
	Weight	
Education	4	
Social services (average of children's and adults)	4	
Environment	2	
Housing	2	
Libraries and leisure	1	
Benefits	1	
Use of resources	1	
3. Maximum and minimum scores are as follows (counties in brackets):		
	Minimum	*Maximum*
Education	4	16
Social services	4	16
Environment	2	8
Housing	2(0)	8(0)
Libraries and leisure	1(0)	4(0)
Benefits	1	4
Use of resources	1	4
Total	15(12)	60(48)
4. Combine core service scores to reach overall judgement:	*Single tier*	*Counties*
1 = lowest	Less than 30	Less than 24
2	30-37	24-29
3	38-45	30-36
4 = highest	More than 45	More than 36
5. Combine score of how council is run from 1 (poor) to 4 (excellent) – weighted average of less than 2 gives overall score of 1 and weighted average of more than 3 gives overall score of 4		
6. Application of rules (to overall category):		
Rule 1: Must score at least 3 (2 stars) on education, social services combined star rating, and financial standing to achieve a category of excellent overall.		
Rule 2: Must score at least 2 (1 star) on education, social services combined star rating, and financial standing to achieve a category of fair or above.		
Rule 3: Must score at least 2 (1 star) on all core services to achieve a category of excellent overall.		

Note: Scores are converted as follows:
Education: 0 star = 1, 1 star = 2, 2 stars = 3, 3 stars = 4
Social services (average score for children and adults): not serving people well = 1, serving some people well = 2, serving most people well = 3, serving people well = 4

Education and social services (which for instance each have a star rating system underpinning them) receive a higher weighting in the overall composite CPA score. In the final step, a set of decision rules are applied (similar to the Trust star ratings in healthcare) which again may impact on the final CPA score given to an authority.

There are significant rewards for high performing councils in the form of:

1. Less ring-fencing
2. Fewer and 'lighter touch' inspections
3. Fewer planning requirements
4. Freedom to use income from fines

The reaction of local government councils to CPA has been mixed with strong support for the self-assessment and peer assessment aspects as well as the financial freedoms, but concern over whether the overall judgements fairly represent performance.

3.2.1 Housing, Environment, Use of Resources and Libraries and Leisure

In the CPA, for earlier years, Best Value Performance Indicators (BVPIs) were scored for the service blocks of Housing, Environment, Use of Resources and Libraries & Leisure. The general approach was to award a percentile ranking to each Performance Indicator (PI) for an authority based on an all-England comparator group.

This was a two-step process, as shown in the following table below, based on ten fictitious authorities.

Step 1
The percentile is calculated relative to the all-England comparator group.

Step 2
The raw PI values are turned into negative figures and the percentile is calculated. These two percentiles are then averaged.

Table 8: The percentile approach for scoring performance indicators

	Raw PI	Step 1	Negative Raw PI	Step 2	Average
Authority 1	29.4	0	-29.4	0.000000	0
Authority 2	41.0	0.222222	-41.0	0.222223	0.22
Authority 3	42.0	0.333333	-42.0	0.444445	0.39
Authority 4	56.3	0.777777	-56.3	0.777778	0.78
Authority 5	38.0	0.111111	-38.0	0.111112	0.11
Authority 6	63.2	1	-63.2	1.000000	1.00
Authority 7	42.0	0.333333	-42.0	0.444445	0.39
Authority 8	45.3	0.555555	-45.3	0.555556	0.56
Authority 9	63.0	0.888888	-63.0	0.888889	0.89
Authority 10	48.8	0.666666	-48.8	0.666667	0.67

Source: Audit Commission (2003a)

In this example, a high value on the raw PI is desirable. The highest performer (Authority 6) gets a percentile of 1 (equivalent to 100th percentile), and the worst, Authority 1, gets a percentile of 0 (i.e. zero percentile).

Within each domain or service block, there is a basket of PIs and the percentile scores for each PI in the block are averaged to give an average percentile for the service block, as shown in the example in the following table for Use of Resources.

Table 9: Averaging the percentiles for the service block: Use of Resources - Financial Administration

	BVPI 8	BVPI 9	BVPI 10
Percentiles for Authority 10	0.49	0.26	0.62
Average percentile for this service block		**0.46**	

Source: Audit Commission (2003a)

This average percentile is scored 1 - 4 (with 4 as the highest score) using the thresholds set out in the following table. These thresholds reflect the fact that it is more difficult to achieve a good performance on every PI in a basket when there are a large number of PIs. Therefore, where an authority has an average percentile equivalent to having three-quarters of its PIs in the top quartile for any service block, and none in the lowest quartile, it will score a 4 and vice versa to score a 1.

Table 10: Converting average percentiles to scores

Average percentile of:	Scores:
above 0.625	4
above 0.5, below or equal to 0.625	3
above 0.375, below or equal to 0.5	2
equal to or below 0.375	1

Source: Audit Commission (2003a)

However, sometimes there are fewer than four PIs within a block in which case thresholds are adjusted as follows:

Table 11: Adjusted thresholds when there are fewer than 4 PIs

Scoring a 1 when there are fewer than 4 PIs		Scoring a 4 when there are fewer than 4 PIs	
Number of PIs	Adjusted threshold	Number of PIs	Adjusted threshold
1	0.25	1	0.75
2	0.31	2	0.69
3	0.33	3	0.67

Source: Audit Commission (2003a)

To take year on year changes into account, percentiles were ‘pegged’ at 2000/01 values. Thus if in 2000/01 a performance of 52% on a PI meant that an authority was awarded the 80th percentile, a performance of 52% in 2001/02 would attract the same, 80th percentile. If the percentiles were not pegged, a performance of 52% in 2001/02 may have attracted a percentile of only 75th, say, if many authorities had improved their performance. Thus, pegging provides an opportunity for all authorities to improve their PI score year on year.

Authorities’ performance was scored against an all-England comparator group in most cases.

Some of the performance indicators were adjusted for social deprivation. The measure of deprivation used was the Average Ward Score of the Index of Multiple Deprivation (IMD) (DETR, 2000). Linear regression analysis was used to examine the relationship between social deprivation and various performance indicators and adjustments were made where a relationship was found to be statistically significant, and considered to be a causal one. User satisfaction PIs were thus adjusted for social deprivation, using the following formula:

$$\textit{Expected satisfaction level} = \textit{constant} + \textit{coefficient} \times \textit{IMD} \qquad (1)$$

where IMD is the Index of Multiple Deprivation of the authority. The expected performance figure is subtracted from the actual performance to give a residual, which is then ranked by percentiles and used in the PI analysis. Using this equation, the coefficient in every case turns out to be negative, reflecting the fact that satisfaction tends to decline as deprivation increases.

In the 2005 CPA some changes were made to the scoring of service blocks Housing, Environment, and Libraries & Leisure (now called Culture). The scores are derived from a performance indicator (PI) score and an inspection score. Both the PI and the inspection element will be given a score of 1 – 4, which will then be combined to give the overall score for each service.

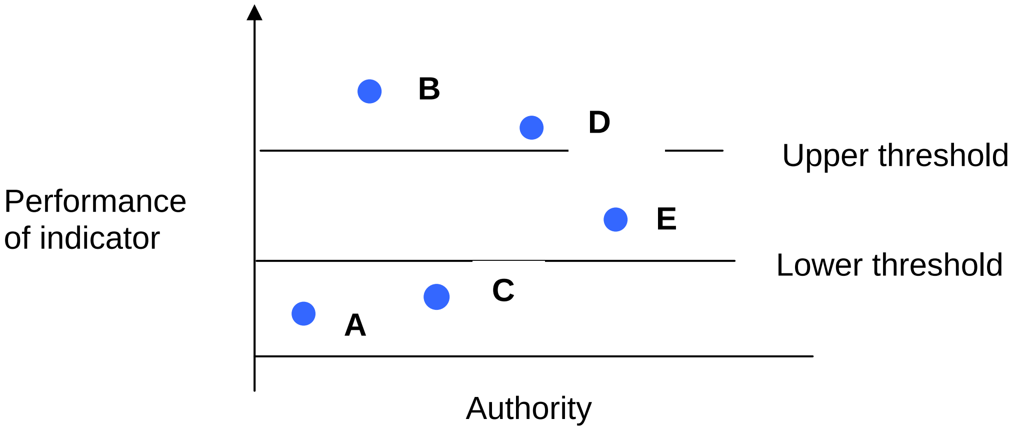
The service assessment scores are determined by applying a threshold to the weighted aggregate PI and inspection element.

Table 12: Thresholds for determining the service assessment score for environment, housing and culture

Weighted average aggregate score	Overall service score
Below 1.85	1
1.85 to less than 2.5	2
2.5 to 3.15	3
Above 3.15	4

Source: Audit Commission (2005) http://www.audit-commission.gov.uk/cpa/downloads/Nov05ServiceAssessmentFrameworksTG.doc

Performance on each PI is compared against pre-determined thresholds. There are two thresholds for each data item – a lower and upper threshold. The following figure shows how the performance of each PI will be compared against the relevant upper and lower thresholds.



For this PI:
Council A scores below the lower threshold
Council B scores above the upper threshold
Council C scores below the lower threshold
Council D scores above the upper threshold
Council E scores between the upper and lower threshold

Figure 1: Performance of PIs against relevant thresholds

Source: Audit Commission (2005) http://www.audit-commission.gov.uk/cpa/downloads/Nov05ServiceAssessmentFrameworksTG.doc

The PI element score is based on the distribution of PIs within the service block, using the following table.

Table 13: Approach to scoring performance indicator data for housing, environment and culture

PI score	Proportion of data items
4	No PIs at or below the lower threshold, and 35% or more PIs at or above the upper threshold
3	No more than 15% of PIs (or 1 PI if 15% equates to less than 1) at or below the lower thresholds, and 25% or more PIs at or above the upper thresholds
2	Any other combination
1	35% or more PIs at or below the lower threshold

Source: Audit Commission (2005) http://www.audit-commission.gov.uk/cpa/downloads/Nov05ServiceAssessmentFrameworksTG.doc

Using this approach means that, so long as minimum standards are broadly being met, the council can determine which areas within the service it wishes to drive performance on towards the upper threshold in order to score higher than a two for the PI element score for the service.

In terms of setting thresholds, where there are national requirements, these are used. Where there is a 'yes/no' response, the yes/no answer determines the upper or lower threshold. In other cases, the default thresholds are linked to quartiles as follows:

- the lower threshold is set at the 25th percentile; and
- the upper threshold is set at the 75th percentile.

Quartiles are calculated on an all-England basis. Thresholds for scoring purposes normally remain unchanged for two years.

Similarly to previous years, adjustments are made to a small number of PIs where a sufficiently large correlation exists between performance and deprivation (for example, surveys of user satisfaction and council tax collection). The Audit Commission revised their approach to such adjustments in 2005. In particular:

- reviewing all PIs to assess the extent of correlation;
- using PIs that do not show such a correlation, in preference to ones that do; and
- where using a PI that does show a correlation, adjusting the PI to reduce the correlation (for example, basing measurement on change in performance rather than absolute performance).

There are still 6 PIs where an adjustment is applied at the level of the individual PI. As in previous years the adjustment is made by uplifting the PI data using the coefficients of a linear regression model using the following equation:

Adjusted PI value = original PI value – linear function of deprivation (2)

A generic example of such a calculation is as follows:

Original PI value = 38%
Gradient = -0.25
Deprivation (IMD 2004 average score) = 36
Adjusted PI = Original PI value – (Gradient x IMD 2004)
Adjusted PI = 38% - (-0.25 x 36)
= 38% - (-9)
= 47%

In the above example the adjustment for deprivation results in the authority's performance being increased by nine percentage points.

3.2.2 Benefits

The Benefit Fraud Inspectorate (BFI), part of the Department for Work and Pensions, undertakes an assessment of the Housing Benefit and Council Tax Benefit service provided by each district council. This is in many cases not a full inspection but an evaluation based on self-assessment. BFI will then use this plus other performance information including BVPIs to produce an overall assessment and report. The BFI report will give an overall rating for both current performance and for capacity to improve. The ratings are based on a five-point scale (poor, fair, fair to good, good, and excellent).

The Audit Commission uses a 4 point scale on the CPA scorecard, hence the BFI's assessment is translated to the Audit Commission's service scorecard using the thresholds shown in the following table.

Table 14: Benefit Fraud Inspectorate assessments within the CPA

BFI assessment	**Audit Commission's scorecard**
Excellent, Good, Fair towards good (80% or above)	4
Fair (60-79%)	3
Fair (40-59%)	2
Poor (0-39%)	1

Source: Benefit Fraud Inspectorate (2003)

3.2.3 Social services

The Social Services Inspectorate (SSI) (now the Commission for Social Care Inspection) produces Personal Social Services (PSS) Star ratings that assess the current performance and prospects for improvement of social services in the areas of services for children and adults (Department of Health, 2002c; Department of Health, 2003; Commission for Social Care Inspection, 2003).

In May 2002, the Department of Health SSI published the first set of social services PSS star ratings. These covered all councils with social services responsibilities in England using all the evidence available at that time. A second updated set of star ratings was published in November 2002, including more up to date performance indicators and inspections, followed by more star ratings published in 2003, 2004 and 2005.

These performance ratings have been formulated from evidence from published Performance Indicators, inspection, Social Services Inspectorate (SSI) / Audit Commission Joint Reviews, self-assessment, and reviews of plans and in-year performance information from both the SSI and the external auditors for each council.

The social services star rating feeds directly into the local government CPA. A council must receive a good star rating for their social services in order to receive the highest comprehensive performance assessment rating. The performance indicators are selected against the following four criteria:

- *Importance* – clearly relating to government priorities;
- *Ease of interpretation* – not requiring further contextual data to understand, with clear criteria to identify good and bad performance;
- *Reliable data* – the data provided by councils are believed to be reliable and definitions of indicators sufficiently capture good practice;
- *Attributable to social services* – the level of the indicator is largely due to the performance of social services, rather than other factors or agencies.

Domains include meeting national priorities, cost and efficiency, effectiveness of service delivery, quality, fair access to services, and prospects for improvement.

Judgements for children and adults services are given. In both cases, a judgement for both current performance and prospects for improvement is given. This results in a total of four judgements underpinning the overall rating. Once the judgements have been reached, a set of decision rules is used to combine them with the weightings to produce a final Star rating.

The principles underlying the decision rules are as follows:

- current performance is weighted more heavily than prospects for improvement;
- adult services and children's services are given equal weight;
- a "failure" in either adult services or children's services will result in zero stars, no matter how good the other services are.

A subset of performance indicators are defined as the Key Performance Indicators and are each given a threshold value determining the maximum judgment that can be given to the indicator. For these, a council could not be judged to be performing well if it failed to reach a specified band of performance. There are minimum standards for both children and adult performance indicators, and a council will have to meet all the minimum standards in order to receive one of the higher judgments.

The following table shows how the Star ratings are presented.

Table 15: Star ratings for social services performance

	Performance rating	Children's services		Adults' services	
		Current performance - Serving people well?	Improvement prospects?	Current performance - Serving people well?	Improvement prospects?
Council 1	-	No	Poor	Most	Promising
Council 2	:	Some	Uncertain	Some	Promising
Council 3	::	Most	Promising	Yes	Uncertain
Council 4	:::	Most	Excellent	Yes	Promising

Source: Department of Health (2002c)

Social services are provided or arranged by local councils, but are often planned and delivered in partnership with the NHS and other council services. The social services star rating is therefore designed to be compatible with performance information for both the NHS and other local government services.

3.2.4 Education

The Office for Standards in Education (OFSTED) is a non-ministerial government department independent of the Department for Education and Skills (DfES). OFSTED produces an 'education profile', or scorecard, with two separate Star Ratings for each local education authority (LEA) in the areas of current performance and improvement (as well as capacity to make further improvements), similar to social services. Five domains are assessed: school improvement, special educational needs, social inclusion, lifelong learning and strategic management of education (OFSTED and DfES, 2002). The league tables produced by the Department for Education and Skills for primary and secondary schools and colleges in the UK contribute to the CPA through the 'education profile'.

All schools are inspected at least once within six years. Inspectors make judgements on a seven-point scale as follows: Excellent 1, very good 2, good 3, satisfactory 4, unsatisfactory 5, poor 6, and very poor 7. The Evaluation Schedule applies to the inspection of all nursery, primary, secondary and special schools, pupil referral units and any other publicly funded provision (Office for Standards in Education, 2003). The Evaluation Schedule covers the following key areas:

1. Effectiveness of the school
2. Standards achieved by pupils
3. Quality of education provided by the school
4. Leadership and management of the school

Most elements of the education profile are made up by combining performance indicators and inspection judgements. The education profile has the potential for fifteen elements or assessments to be made (across the three areas - current performance, improvement, capacity; and five domains - school improvement, special educational needs, social inclusion, lifelong learning, strategic management). Each of the elements across the five domains is then aggregated to give an overall assessment score for each of the three areas.

There are 45 indicators used to feed into the elements of the profile. Of these, 24 are performance indicators and 21 are LEA inspection judgements. The performance indicators show how well an LEA is doing compared to all LEAs. The inspection judgements show how well the LEA is doing compared to the standards set in OFSTED's criteria for inspection judgements.

No adjustment is made to the profile for social deprivation. This is for two main reasons. Firstly, OFSTED argue that national funding is designed to recognise the challenges faced by an LEA. Secondly, nearly half (21 out of 45) of the indicators used are based on inspection judgements, which are made taking the context of an LEA into account.

Each indicator is converted to a categorical score on a five-point scale with 1 being the highest and 5 being the lowest score. For the performance indicators, all LEAs are ranked and the score is then determined by allocating the top 10% a 1, the next 20% a 2, the next 40% a 3, the next 20% a 4 and the remaining 10% a 5. For inspection judgements, which are made in the first instance on a seven-point scale, the inspection grades are converted to a five-point scale.
The scores within each of the fifteen elements are then added together and divided by the number of scores to give an overall score for the element. This is the score shown on the education score card for the element.

A few weights have been introduced into the creation of the profile with respect to the construction of each element and the weights attached to certain indicators and inspection judgements. Otherwise, there are no other weightings in the profile. Each element has an equal effect on the overall score for each area.

Table 16: Operation of performance and improvement rules

Performance stars	Improvement score	Indicated improvement	Improvement stars	Capacity score	Indicated capacity	Improvement stars
:::	1.0 - 2.9	Proven	:::	-	-	-
:::	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	::
:::	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	:
::	1.0 - 2.9	Proven	:::	-	-	-
::	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	::
::	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	:
:	1.0 - 2.9	Proven	-	1.0 - 2.9	Secure	::
:	1.0 - 2.9	Proven	-	3.0 - 5.0	Not secure	:
:	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	:
:	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	-
-	1.0 - 2.9	Proven	-	1.0 - 2.9	Secure	::
-	1.0 - 2.9	Proven	-	3.0 - 5.0	-	:
-	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	:
-	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	-

Source: Office for Standards in Education (2002a)

The Audit Commission model uses scores on a four-point scale to feed into the CPA for a council. To this end, the average score, and the ranked position, of an LEA are used to determine its category on the four-point scale.

The ranked list for current performance is allocated Star ratings on the basis of the inspection grades and performance quotas for each category. The improvement categories are derived differently and the rules used to allocate Star ratings for improvement are shown in table 16.

The following table shows an example of a LEA education profile, for the city of York.

Table 17: Example of Local Education Authority CPA education scorecard (for the City of York)

Aspect	Current performance	Indications of improvement	Capacity to make further improvement
School Improvement	2.1	2.4	1.6
SEN	2.5	1.0	4.0
Social Inclusion	1.6	2.8	2.3
Life Long Learning	2.7	2.0	2.0
Strategic Management	1.3	-	2.5
Average Score	2.0	2.3	2.2
Category	:::	:::	

Note: The Average Score thresholds for the Performance star ratings are as follows:
3 star is obtained if the Average Score is less than or equal to 2.38
2 star is obtained if the Average Score is less than or equal to 3.34 but more than 2.38
1 star is obtained if the Average Score is less than or equal to 3.75 but more than 3.35
- star is obtained if the Average Score is greater than 3.75

Source: Office for Standards in Education (2002b)

The whole CPA process is therefore hugely complex and built on a structure of underlying composite indices, each with a huge number of methodological choices underpinning them. Some of the potential pitfalls with the process include the application of decision rules to construct the star ratings, the widespread use of categorical variables with potentially arbitrary thresholds, and the application of percentile thresholds, the opaque use of weights, and the inconsistency in dealing with exogenous factors. One of the main differences though with the Star rating system applied to health care is the reward schedule attached to the CPA. Authorities have to deal with more inspections if they perform poorly, whereas in NHS hospital Trusts, the management team can effectively be replaced.

4. Empirical analysis

This section explores some of the technical issues involved in the construction of a composite indicator. We use data to construct a new generic composite index. Each of the steps of constructing a new composite are then simulated and the robustness of the scores and rankings of individual organisations are examined.

4.1 The data

We used longitudinal panel data on two key composite performance measures used in 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 from 2001/02 to 2004/05 covering around 110 Best Value Performance Indicators (BVPIs).

The purpose of using this data is not to reproduce the Star and CPA ratings, but rather to use the two datasets for illustrative purposes. While there are limitations to the coverage, consistency and accuracy of the data, we had limited time available in this study to explore all of these issues. Since the data is used in its existing form for regulation, the short duration of the project meant we were restricted to the use of this publicly available data. However, we devoted considerable effort to linking it over time, collating it for our analysis and getting it into a useable format. The remaining data limitations do not affect the messages arising from our analysis.

4.2 Methods and results

Our research methodology was to construct a generic composite indicator for each of the two sectors, using them to examine the sensitivity of rankings of organisations to the various methodological issues involved at each step in the construction. The purpose was not to replicate the ratings that currently exist in these sectors, but rather to explore the sensitivity of a generic composite to the various methodological judgements made at each step.

4.2.1 Choosing the indicators

In order to make this generic measure practical, but also realistic, we selected the underlying indicators for each composite applying the following criteria:

- The indicators should cover the broad range of performance (we used factor analysis to obtain the key dimensions and explored correlations between indicators);
- The indicators should not have large numbers of missing values;
- They should exhibit stable statistical properties (we used tests of skewness, kurtosis and normality);
- The indicators should preferably be available for more than 1 year.

In order to explore the relationship between different indicators we used factor analysis. Factor analysis is a data reduction method with the principal idea being that one can describe a set of *p* variables in terms of a smaller number of *m* factors. Thus each of the variables takes the form $X_i = a_i F + e_i$ where X_i is the i^{th} standardised score with a mean of zero and a standard deviation of one, a_i is a constant, *F* is a ‘factor’ value with a mean of zero and a standard deviation of one and e_i is the part of X_i that is specific to the i^{th} score only. A general form of this model is as follows:

$$\begin{aligned} X_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \ldots + \alpha_{1m}F_m + e_1 \\ X_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \ldots + \alpha_{2m}F_m + e_2 \\ &\vdots \\ X_p &= \alpha_{p1}F_1 + \alpha_{p2}F_2 + \ldots + \alpha_{pm}F_m + e_p \end{aligned} \qquad (3)$$

where X_i is a variable with a mean of zero and a standard deviation of one, $\alpha_{i1}, \alpha_{i2}, \ldots \alpha_{im}$ are the factor loadings related to the variable X_i, $F_1, F_2, \ldots, F_m$ are *m* uncorrelated common factors, each with a mean of zero and a standard deviation of one and e_i is the specific factor related only to the variable X_i, which has zero mean and is uncorrelated with any other common factor and the specific factors (Joint Research Centre, 2002).

We used rotated factor analysis which produces results which can more readily be interpreted. The purpose of rotated factor analysis is to obtain a clear pattern of loadings, that is, factors that are clearly marked by high loadings for some variables and low loadings for others. Thus a rotation is sought that maximizes the variance and produces a pattern of loadings on each factor that is as diverse as possible, lending itself to easier interpretation.

This exercise was exploratory, to examine the associations between indicators.

Results for the healthcare data showed the following set of independent factors which corresponded to the correlations found. These were:

- Inpatient satisfaction
- Waiting times
- Cancelled operations
- Readmission rates
- Cancer waits
- Outpatient waits
- Financial balance

Rotated factor analysis was performed by service domain for local government. Results are too complex to reproduce in this report, but again corresponded to underlying correlations found.

The correlation matrix for all performance indicators was produced but is too complex to include in the report in its entirety.

We then explored the descriptive statistics for the raw performance indicators in each sector. These included the number of observations (any missing data and availability over time), the mean, median and standard deviation as well as measures of skewness and kurtosis which give an indication of the type of distribution of the variable.

We used the factor analysis, correlations and descriptive statistics to select a smaller number of indicators from each dataset.

4.2.2 The chosen indicators

The analysis above resulted in a subset of 10 performance indicators for healthcare in 2001/02 and a subset of 35 performance indicators for local government in 2003/04. The following table describes the performance indicators from the healthcare dataset.

Table 18: Variable names and descriptions, healthcare

Performance indicator	Indicator variable	Variable name
Clinical focus		
Death within 30 days of surgery (non-elective admissions)	Deaths within 30 days of surgery for non-elective admissions to hospital, per 100,000 patients (age and sex standardised, includes deaths in hospital and after discharge)	d_esurgstd
Emergency readmission to hospital following treatment for a fractured hip	Emergency readmissions to hospital within 28 days of discharge following treatment for a fractured hip, as a percentage of live hip fracture discharges (age and sex standardised)	readm_hip
Returning home following hospital treatment for fractured hip	Percentage of patients discharged back to usual place of residence within 28 days of emergency admission to hospital with a hip fracture, all ages (age and sex standardised)	dis_hippc
Patient focus		
Six month inpatient waits	Percentage of patients waiting less than 6 months for an inpatient admission	wait6pc
Thirteen week outpatient waits	Percentage of patients seen within 13 weeks of GP written referral for first outpatient appointment	outwt13wkpc
Inpatient survey - coordination of care	Combined score of questions around organisation of emergency care, organisation of admissions process, staff giving conflicting information and members of staff taking overall charge of care	inp_survey_coord
Capacity and capability focus		
Data quality	Summary measure of Hospital Episode Statistics (HES) data quality for NHS trusts with in-patient activity	dqi_pc
Staff opinion survey	Responses from NHS-employed staff opinion survey on satisfaction with employer	staff_survey
Junior doctors' hours	Percentage of Junior Doctors complying in full with the New Deal on Junior Doctors' Hours	jundocpc
Sickness absence rate	Amount of time lost through absences as a percentage of staff time available for directly employed NHS staff	sick_rate

The descriptive statistics for these variables are shown in table 19.

Table 19: Descriptive statistics of variables used to form composite, healthcare

Variable	n	mean	std.dev	min	max
Healthcare					
Patient outcomes					
Death within 30 days of surgery per 100,000 patients (non-elective admissions)	149	2951	528.3	1438.7	4386.5
Emergency readmission to hospital within 28 days of discharge following hip fracture as percent of live hip fracture discharges	149	8.0	2.4	1.7	14.7
Returning home within 28 days of emergency admission to hospital for fractured hip	148	47.9	8.3	28.3	75.1
Access					
Percent patients waiting 6 months or less for an inpatient admission	180	78.8	8.7	60.8	100.0
Percent patients seen within 13 weeks of GP referral for first outpatient appointment	175	74.6	9.2	49.9	97.7
Satisfaction surveys					
Inpatient survey satisfaction with coordination of care	171	68.0	4.2	56.7	79.7
Responses from NHS-employed staff survey on satisfaction with employer (0-5-point scale)	168	3.2	0.2	2.7	3.7
Data					
Summary measure of Hospital Episode Statistics (HES) data quality	180	91.0	7.3	66.8	99.2
Staffing					
Percent Junior Doctors complying with New Deal on Working Hours	171	59.4	21.2	0	100
Sickness absence rate - Amount of time lost through absence as a percent of staff time available	168	4.6	0.7	2.5	6.7

The following table describes the variables chosen in the local government dataset.

Table 20: Variable names and descriptions, local government

Performance indicator	Indicator variable	Variable name
Benefits		
Renewal claims on time BV78c	Percent renewal claims processed on time	v1686st
Housing Benefit Security BV76c	Housing Benefit Security – Number of fraud investigations per 1000 caseload	v5624st
Community Safety		
Burglaries per 1000 households BV126a	Number of recorded domestic burglaries per 1000 households	mv1704st
Racial incidents further action BV175	Percent racial incidents that resulted in further action	v1712st
Corporate Health		
Community Strategy BV1a	Community Strategy developed in collaboration with local strategic partnership?	v1641st
Senior women BV11a	Percent of top 5% of earners in authority that are women	v1649st
Ill health retirements / staff BV15	Percent employees retiring on grounds of ill health as percent of total staff	mv1653st
Working age (18-65) people with disabilities BV16b	Percent working age (18-65) people with disabilities as percent of total economically active	v1655st
Working age (18-65) people from ethnic minorities BV17b	Percent working age (18-65) people from ethnic minorities as percent of economically active minority ethnic community	v1657st
Types of interactions delivered electronically BV157	Percent of interactions with citizens delivered electronically	v1659st
Culture and Libraries		
Score on Creating Opportunities checklist BV114	Score on Creating Opportunities checklist for adoption of Local Cultural Strategy	v1698st
Visits to libraries BV117	Number of visits to libraries per 1000 population	v1700st
Education		
Secondary schools 25% + surplus places BV34b	Percent secondary schools with 25% or more places unfilled and at least 30 surplus places	mv1748st
Pupils 5 or more GCSEs: A*-C BV38	Percent 15 year old pupils with 5 or more GCSEs at grades A*-C	v1752st
SENs in 18 weeks with exceptions BV43b	Percent statements of Special Educational Need (SENs) issued in 18 weeks with exceptions	v1757st
Unauthorised absences secondary schools BV45	Percent half days missed due to unauthorised total absences in secondary schools	mv1759st
Schools subject to special measures BV48	Percent schools identified by OFSTED as requiring special measures	v1761st
Perm. excluded pupils alternative tuition: 20 hours BV159d	Percent permanently excluded pupils attending alternative tuition: 20 hours or more a week	v1765st
Environment		
Recycling BV82a	Percent total tonnage of household waste which has been recycled	v1715st
Composting BV82b	Percent total tonnage of household waste which has been sent for composting	v1716st
Household waste collected BV84	Number of kilograms of household waste collected per head	v1719st
Housing		
Priv. dwellings 6 months empty: returned to occupation BV64	Number of private dwellings 6 months empty: returned to occupation or demolished	v1665st
Length of stay in hostels BV183b	Average length of stay of households in hostels	v1679st
Planning		
New homes on brown field sites BV106	Percent new homes built on brown field sites	v1689st
Major planning apps in 13 weeks BV109a	Percent major planning applications in 13 weeks	v1691st
Standard searches 10 working days BV179	Percent standard searches carried out in 10 working days	v1694st
Social Services		
Children with 3 + placements BV49	Percent children looked after with 3 or more placements	mv1769st
Cost intensive care for adults BV52	Average gross weekly cost of intensive social care for adults and older people	mv1772st
Older people helped to live at home BV54	Older people helped to live at home per 1000 population aged 65 or over	v1774st
Needs statements BV58	Percent people receiving needs statement and how they will be met	v1777st

Care leavers in education / training / employment BV161	Percent care leavers in education / training / employment	v1778st
Transport		
Condition classified non-principal roads BV97a	Condition classified non-principal roads by Coarse Visual Inspection (CVI) survey	v1725st
Road accidents: car users BV99d	Number of road accident casualties per 100,000 population: car users	mv1733st
Footpaths easy to use BV178	Percent total length of footpaths easy to use by the public (Country Agency and CSS Survey)	v1740st
Principal roads not needing major repair BV186a	Percent principal roads not needing major repair / average structural expenditure per km	v1742st

The descriptive statistics for these variables are shown in table 21.

Table 21: Descriptive statistics of variables used to form composite, local government

Variable	n	mean	std.dev	min	max
Local government					
Corporate Health					
Community Strategy developed in collaboration with local strategic partnership?	150	0.8	0.4	0	1
Percent of top 5% of earners in authority that are women	150	39.6	7.3	0	53.6
Percent employees retiring on grounds of ill health as percent of total staff	150	0.4	0.2	0	1.2
Percent working age (18-65) people with disabilities as percent of total economically active	150	14.3	4.0	0	23.4
Percent working age (18-65) people from ethnic minorities as percent of economically active minority ethnic community	150	10.3	12.1	0	59.3
Percent of interactions with citizens delivered electronically	150	66.4	17.3	15.8	100
Education					
Percent secondary schools with 25% or more places unfilled and at least 30 surplus places	148	7.1	7.1	0	33.3
Percent 15 year old pupils with 5 or more GCSEs at grades A*-C	149	50.2	8.1	32.1	87.5
Percent statements of Special Educational Need (SENs) issued in 18 weeks with exceptions	150	67.5	22.7	5.6	100
Percent half days missed due to unauthorised total absences in secondary schools	148	8.4	1.1	2.1	11.4
Percent schools identified by OFSTED as requiring special measures	149	1.4	1.4	0	7
Percent permanently excluded pupils attending alternative tuition: 20 hours or more a week	148	68.7	25.3	0	100
Social Services					
Percent children looked after with 3 or more placements	149	12.6	3.6	0	21
Average gross weekly cost of intensive social care for adults and older people	150	459.0	80.1	306	735
Older people helped to live at home per 1000 population aged 65 or over	150	89.0	26.1	35	183
Percent people receiving needs statement and how they will be met	150	88.8	11.9	17.8	100
Percent care leavers in education / training / employment	150	56.6	14.9	0	100
Housing					
Number of private dwellings 6 months empty: returned to occupation or demolished	114	146.8	227.7	0	1319
Average length of stay of households in hostels	115	17.1	33.4	0	238
Benefits					
Percent renewal claims processed on time	116	62.6	19.5	0	98.5
Housing Benefit Security – Number of fraud investigations per 1000 caseload	116	32.1	21.9	0	127.9
Environment					
Percent total tonnage of household waste which has been recycled	149	11.8	3.7	3.8	20.6
Percent total tonnage of household waste which has been sent for composting	148	4.5	3.5	0	21.3
Number of kilograms of household waste collected per head	149	494.2	56.3	319.2	659

Transport					
Condition classified non-principal roads by Coarse Visual Inspection (CVI) survey	148	19.7	11.0	0	58
Number of road accident casualties per 100,000 population: car users	148	27.8	14.2	5.3	87
Percent total length of footpaths easy to use by the public (Country Agency and CSS Survey)	135	69.4	21.6	5	100
Percent principal roads not needing major repair / average structural expenditure per km	148	72.4	46.3	0.1	264
Planning					
Percent new homes built on brown field sites	115	82.0	21.8	17.9	100
Percent major planning applications in 13 weeks	150	52.1	15.8	13.0	100
Percent standard searches carried out in 10 working days	116	93.8	14.0	16.1	100
Culture and Libraries					
Score on Creating Opportunities checklist for adoption of Local Cultural Strategy	149	86.8	25.5	0	100
Number of visits to libraries per 1000 population	149	6378.7	7883.7	5	99731
Community Safety					
Number of recorded domestic burglaries per 1000 households	149	20.0	9.9	6.5	64.1
Percent racial incidents that resulted in further action	145	89.7	22.7	0	100

All indicators were transformed to “more is better” e.g. death rates were converted to survival rates.

We then standardised the performance indicators to a z score (with mean zero, unit variance). Since all indicators are approximately normal and symmetric (one of our criteria for statistical properties), we could assume they are drawn from the distribution ~N(0,1). Hospitals and local authorities with missing data were excluded from the analysis, giving a final sample of 117 hospitals and 97 local authorities.

In the first instance, to create a composite score for each hospital and local authority, we aggregated the transformed variables in a linear fashion with a uniform weight and created a ranking based on the composite score.

The following tables give the descriptive statistics for the transformed variables for healthcare and local government respectively. Since the raw data was not first re-scaled, the indicators will not have exactly the same range (max – min). The transformed variables also mostly appear to be approximately normal and symmetrical.

Skewness is a measure of the lack of symmetry of a distribution. If the coefficient of skewness is zero, the distribution is symmetric. If the coefficient is negative, the median is usually greater than the mean and the distribution is skewed left. If the coefficient is positive, the median is usually less than the mean and the distribution is skewed right. Kurtosis is a measure of peakedness of a distribution. The smaller the coefficient of kurtosis, the flatter the distribution. The normal distribution has a coefficient of kurtosis of 3 and provides a convenient benchmark. The test for normality based on D’Agostina *et al* (1990), runs separate tests for normality based on skewness and kurtosis and then combines the two tests into an overall Chi-squared test statistic. Most of the variables have a Prob > Chi-squared > 0.05 which suggests they are not significantly skewed, although this was less the case for the local authority data.

The descriptive statistics for the new composite indicator are also shown in the tables. As expected, the standard deviation for the composite will be much larger than for each of the underlying indicators. The hospitals and local authorities could then also be ranked on the new composite.

Table 22: Descriptive statistics of standardised variables and new composite, healthcare

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
mind_esurgstdst	117	0	-0.021	1	-2.669	2.068	-0.009	2.520	0.516
dis_hippcst	117	0	-0.053	1	-2.284	3.200	0.322	3.180	0.258
wait6pcst	117	0	-0.012	1	-2.257	2.637	0.231	3.011	0.530
outwt13wkpcst	117	0	-0.009	1	-2.414	2.739	-0.011	2.888	0.998
inp_survey_coordst	117	0	0.043	1	-3.009	2.494	-0.162	2.892	0.751
jundocpcst	117	0	0.073	1	-2.398	2.157	-0.287	2.650	0.326
minsick_ratest	117	0	0.089	1	-2.739	2.744	-0.306	3.144	0.301
minreadm_hipst	117	0	0.143	1	-2.697	2.551	-0.422	3.104	0.135
staff_surveyst	117	0	-0.089	1	-3.878	2.640	-0.669	4.576	0.002
dqi_pcst	117	0	0.298	1	-3.578	1.205	-1.505	5.099	0.000
composite	117	0	0.634	3.361	-9.679	7.503	-0.654	3.269	0.019

Table 23: Descriptive statistics of standardised variables and new composite, local government

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
v1686st	97	0	0.116	1	-2.380	1.839	-0.253	2.331	0.1192
v5624st	97	0	-0.115	1	-1.498	4.282	1.798	8.050	0.0000
mv1704st	97	0	0.102	1	-4.070	1.469	-1.056	4.721	0.0002
v1712st	97	0	0.428	1	-5.286	0.428	-3.385	15.666	0.0000
v1641st	97	0	0.318	1	-3.111	0.318	-2.807	8.880	0.0000
v1649st	97	0	-0.061	1	-2.532	2.199	0.021	2.596	0.7555
mv1653st	97	0	0.206	1	-3.451	1.722	-1.139	4.559	0.0001
v1655st	97	0	-0.100	1	-2.534	2.111	0.034	2.355	0.2552
v1657st	97	0	-0.421	1	-0.874	3.076	1.412	3.944	0.0000
v1659st	97	0	0.053	1	-2.485	1.977	-0.169	2.811	0.7634
v1698st	97	0	0.523	1	-3.323	0.523	-2.231	7.233	0.0000
v1700st	97	0	-0.186	1	-3.708	2.357	-0.012	3.992	0.1607
mv1748st	97	0	0.192	1	-3.361	0.954	-0.937	3.345	0.0037
v1752st	97	0	-0.108	1	-2.126	2.304	0.178	2.435	0.3129
v1757st	97	0	0.142	1	-2.551	1.397	-0.604	2.539	0.0420
mv1759st	97	0	0.034	1	-2.310	5.198	0.973	9.361	0.0000
v1761st	97	0	-0.157	1	-0.925	3.414	1.042	3.749	0.0009
v1765st	97	0	0.088	1	-2.687	1.168	-0.749	2.925	0.0209
v1715st	97	0	-0.087	1	-2.093	2.788	0.481	3.204	0.1044
v1716st	97	0	-0.190	1	-1.227	5.365	2.001	10.457	0.0000
v1719st	97	0	0.037	1	-2.457	2.527	0.022	2.754	0.9598
v1665st	97	0	-0.339	1	-0.616	5.486	3.239	14.771	0.0000
v1679st	97	0	-0.299	1	-0.730	5.214	2.414	10.635	0.0000
v1689st	97	0	0.315	1	-2.729	0.941	-1.089	3.285	0.0012
v1691st	97	0	0.147	1	-2.533	2.210	-0.266	2.643	0.4374
v1694st	97	0	0.437	1	-6.504	0.447	-3.781	21.555	0.0000
mv1769st	97	0	-0.185	1	-2.282	3.168	0.395	3.527	0.1096
mv1772st	97	0	0.204	1	-3.545	1.852	-1.030	4.359	0.0004
v1774st	97	0	-0.117	1	-1.904	3.869	0.714	4.160	0.0056
v1777st	97	0	0.208	1	-5.058	1.066	-2.209	9.929	0.0000
v1778st	97	0	0.028	1	-2.565	2.153	-0.179	2.671	0.6529
v1725st	97	0	-0.144	1	-1.672	3.597	1.090	4.412	0.0002
mv1733st	97	0	0.213	1	-3.496	1.604	-1.119	4.232	0.0003
v1740st	97	0	0.115	1	-2.840	1.223	-0.946	3.595	0.0024
v1742st	97	0	-0.320	1	-1.417	4.246	1.578	6.253	0.0000
composite	97	0	-0.629	5.925	-21.476	12.068	-0.183	3.623	0.2622

4.2.3 Examining uncertainty

In order to explore the degree to which composite measures provide a robust summary of performance, we used Monte Carlo simulations, a stochastic modelling approach, to explore the impact of uncertainty on the composite measure. This enables us to examine the technical decisions made at each step in the construction of the composite and the impact these choices have on the robustness of the rankings.

Monte Carlo simulation is a stochastic technique used to solve mathematical problems. The term ‘stochastic’ means that it uses random numbers and probability statistics to obtain an answer. Simulation means it is an analytical method meant to imitate a real-life system which is used when other analyses are too mathematically complex or too difficult to reproduce (Mooney, 1997).

Without the aid of simulation, the construction of the composite index will only reveal a single outcome. Monte Carlo simulation randomly generates values for uncertain variables over and over to simulate a model.

The term "Monte Carlo" comes from the name of the city in Monaco where the city's main attractions are casinos, which run games of chance such as roulette wheels, dice and slot machines; games which exploit the random behaviour of each game. The random behaviour in games of chance is similar to how Monte Carlo simulation selects variable values at random to simulate a model. For each uncertain variable (one that has a range of possible values), the possible values are defined with a probability distribution. The type of distribution selected is based on the conditions surrounding that variable. Distribution types include normal, log normal and so on. In this example, the variables all have a z score distribution ~N(0,1).

A simulation calculates multiple repetitions of a model by repeatedly sampling values from the probability distributions for the uncertain variables. Simulations can consist of as many repetitions as chosen. During a single repetition, a value is randomly drawn from the defined possibilities (the range and shape of the distribution) for each uncertain variable and selected to fit a probability distribution. This random draw process is then repeated many times. Each time a value is randomly drawn, it forms one possible solution (or in this case composite indicator). Together, these repetitions give a range of possible solutions, some of which are more probable and some less probable. Accuracy of this solution can be improved by increasing the number of repetitions.

This exercise uses 1000 repetitions and thus produces 1000 composite indices which produce a range of outcomes. These can then be used to produce 95 percent uncertainty intervals around the composite (truncating the data at the 2.5 and 97.5 percentiles). The interpretation of these intervals is that on 1000 repetitions, there is a 95 percent probability that the composite index will fall within the interval presented. We can thus show the range of composite scores a single organisation obtains over 1000 simulations.

Simulations were performed by drawing samples from a multivariate normal distribution to emulate the 10 indicators in healthcare and the 35 indicators in local government respectively and the correlations of the underlying indicators in each sector with one another. Thus by sampling values from the probability distributions for the performance indicators with the same zero means and the same covariance matrix, 1000 random versions of the above two datasets could be reproduced. The datasets were generated for data with mean zero and variance one ~ N(0,1) for each sample (in each sector). Each sample was drawn with the above correlation structure (of the underlying standardised variables), for 1000 replications. For each sample a new composite was constructed and a ranking based on the new composite.

The following table shows the descriptive statistics for the two simulation samples. The indicators have approximately mean zero unit variance. However the range is a bit wider than for the original composite since there is greater uncertainty around the indicators and across the 1000 draws. Once again the indicators are on average approximately normal.

Table 24: Descriptive statistics of standardised variables from simulations (1000 replications), healthcare

Variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
mind_esurgstdst	117000	-0.0012	-0.0026	1.0011	-4.6569	4.7106	0.619	0.554	0.7416
dis_hippcst	117000	0.0026	0.0042	1.0015	-4.3791	4.2569	0.360	0.629	0.5854
wait6pcst	117000	-0.0010	0.0000	0.9984	-4.1414	4.3906	0.812	0.864	0.9581
outwt13wkpcst	117000	-0.0035	-0.0008	1.0021	-3.9339	4.4132	0.179	0.576	0.3475
inp_survey_coordst	117000	-0.0032	-0.0021	0.9987	-4.5566	4.1575	0.353	0.404	0.4589
Jundocpcst	117000	-0.0029	-0.0033	0.9967	-4.5327	4.4441	0.444	0.037	0.0856
minsick_ratest	117000	0.0027	0.0035	0.9999	-4.3413	4.2740	0.448	0.880	0.7408
minreadm_hipst	117000	0.0050	0.0041	1.0007	-5.5077	4.8468	0.074	0.038	0.0235
staff_surveyst	117000	0.0026	0.0033	0.9991	-4.5720	4.2472	0.972	0.452	0.7524
dqi_pcst	117000	-0.0023	-0.0015	0.9989	-4.4641	4.4311	0.498	0.328	0.4926

Table 25: Descriptive statistics of standardised variables from simulations (1000 replications), local government

Variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
v1686st	97000	0.0003	0.0033	0.9987	-4.1128	4.4526	-0.0088	2.9869	0.3829
v5624st	97000	0.0053	0.0038	0.9957	-4.1505	4.3864	-0.0043	3.0225	0.3113
mv1704st	97000	-0.0030	-0.0013	1.0010	-3.9239	4.3244	0.0063	2.9886	0.5581
v1712st	97000	0.0005	0.0000	0.9966	-4.1470	4.1996	0.0063	2.9852	0.4668
v1641st	97000	-0.0047	-0.0025	1.0010	-4.4432	4.3250	-0.0076	2.9886	0.4844
v1649st	97000	0.0006	0.0029	0.9972	-4.6566	4.7464	0.0009	2.9898	0.8098
mv1653st	97000	-0.0012	-0.0005	1.0017	-5.4827	4.1855	-0.0003	3.0100	0.8125
v1655st	97000	0.0029	0.0021	1.0019	-4.5247	4.9728	0.0038	2.9944	0.8382
v1657st	97000	0.0024	0.0004	1.0006	-5.0967	4.3581	0.0021	3.0147	0.6228
v1659st	97000	-0.0020	-0.0007	1.0033	-5.2464	4.4076	-0.0037	2.9956	0.8666
v1698st	97000	-0.0021	-0.0039	1.0024	-4.3848	4.3877	0.0133	3.0048	0.2299
v1700st	97000	-0.0007	-0.0006	0.9979	-3.9461	4.2839	-0.0009	2.9894	0.7970
mv1748st	97000	0.0012	-0.0024	0.9988	-3.9874	4.4955	0.0068	3.0102	0.5571
v1752st	97000	-0.0028	0.0011	1.0047	-4.3361	4.2834	-0.0091	2.9833	0.2939
v1757st	97000	-0.0016	-0.0012	0.9972	-4.6490	4.6829	0.0094	3.0175	0.2618
mv1759st	97000	-0.0019	0.0009	1.0025	-4.0473	3.9687	0.0005	2.9771	0.3450
v1761st	97000	0.0046	0.0074	1.0003	-4.1907	4.2712	-0.0055	2.9804	0.3612
v1765st	97000	-0.0044	-0.0044	0.9993	-4.7051	5.0809	0.0031	3.0218	0.3546
v1715st	97000	-0.0004	-0.0012	1.0023	-4.2696	4.2264	0.0029	3.0120	0.6961
v1716st	97000	0.0048	0.0022	0.9983	-4.3722	4.2092	-0.0046	2.9945	0.7948
v1719st	97000	0.0028	0.0061	0.9991	-4.4206	4.6479	-0.0040	2.9857	0.5825
v1665st	97000	-0.0021	-0.0038	1.0031	-4.3287	4.5879	-0.0027	2.9990	0.9407
v1679st	97000	0.0001	0.0000	0.9980	-4.3363	5.0054	0.0032	3.0190	0.4430
v1689st	97000	-0.0019	-0.0008	1.0018	-4.0965	4.8233	-0.0139	3.0174	0.1148
v1691st	97000	-0.0022	0.0008	0.9972	-4.2904	4.8216	-0.0050	3.0146	0.5291
v1694st	97000	0.0036	0.0069	0.9974	-4.2216	4.5467	-0.0077	2.9995	0.6199
mv1769st	97000	0.0035	-0.0021	0.9982	-4.3493	4.6754	0.0084	2.9736	0.1373
mv1772st	97000	0.0030	0.0043	1.0013	-4.3741	4.8325	0.0172	3.0071	0.0819
v1774st	97000	0.0002	0.0041	0.9991	-4.4465	4.4441	0.0013	3.0247	0.2917
v1777st	97000	0.0010	-0.0008	1.0010	-4.1366	4.4813	0.0070	2.9976	0.6689
v1778st	97000	0.0012	0.0048	0.9993	-4.6558	4.8914	-0.0101	3.0032	0.4295
v1725st	97000	0.0010	-0.0002	0.9970	-4.1623	4.3281	0.0027	3.0376	0.0583
mv1733st	97000	0.0007	-0.0035	0.9999	-4.3129	4.2899	0.0062	2.9862	0.5043
v1740st	97000	-0.0035	-0.0037	0.9988	-4.7587	4.0123	-0.0096	2.9881	0.3610
v1742st	97000	0.0023	0.0026	0.9993	-4.6308	4.4017	0.0012	2.9815	0.4939

In order to construct the new composite, the original scores from each performance indicator for each hospital and local authority are then added to the simulated dataset to obtain the following descriptive statistics. As can be seen the standard deviation for each indicator increases to approximately 1.4 and the range increases (for the same reasons) from ±4.5 to ±6 in both sectors. The standard deviation on the composite is also commensurately larger at 4.8 for healthcare and 8.3 for local government.

Table 26: Descriptive statistics of standardised variables and new composite from simulations (1000 replications), healthcare

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
mind_esurgstdst	117000	-0.0012	-0.0080	1.4138	-5.9369	5.6063	0.904	0.000	0.0000
dis_hippcst	117000	0.0026	-0.0318	1.4157	-5.6570	7.0279	0.000	0.000	0.0000
wait6pcst	117000	-0.0010	-0.0134	1.4136	-5.7476	5.8587	0.000	0.351	0.0000
outwt13wkpcst	117000	-0.0035	0.0043	1.4125	-5.6276	6.0684	0.506	0.047	0.1122
inp_survey_coordst	117000	-0.0032	0.0149	1.4068	-6.4436	5.4590	0.000	0.000	0.0000
jundocpcst	117000	-0.0029	0.0327	1.4072	-5.6347	6.3681	0.000	0.000	0.0000
minsick_ratest	117000	0.0027	0.0346	1.4105	-5.8227	5.7476	0.000	0.279	0.0000
minreadm_hipst	117000	0.0050	0.0499	1.4143	-6.2602	5.6764	0.000	0.009	0.0000
staff_surveyst	117000	0.0026	0.0409	1.4130	-7.0770	5.9946	0.000	0.000	
dqi_pcst	117000	-0.0023	0.1227	1.4118	-6.7770	5.4617	0.000	0.000	
composite	117000	-0.0013	0.1971	4.7504	-21.5216	17.6654	0.000	0.000	0.0000

Table 27: Descriptive statistics of standardised variables and new composite from simulations (1000 replications), local government

Variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
v1686st	97000	0.0003	0.0299	1.4067	-5.3120	5.7903	-0.0874	2.8441	0.0000
v5624st	97000	0.0053	-0.0891	1.4029	-4.8048	7.4559	0.6243	4.2312	
mv1704st	97000	-0.0030	0.0687	1.4092	-6.7975	5.1957	-0.3743	3.4237	
v1712st	97000	0.0005	0.1576	1.4074	-8.2981	4.6276	-1.1945	6.1748	
v1641st	97000	-0.0047	0.1847	1.4110	-6.4979	4.0977	-0.9869	4.4518	
v1649st	97000	0.0006	-0.0077	1.4083	-6.5174	6.8565	0.0182	2.9136	0.0000
mv1653st	97000	-0.0012	0.0915	1.4141	-6.3924	5.6540	-0.4088	3.4056	
v1655st	97000	0.0029	-0.0096	1.4149	-5.7053	5.7800	0.0107	2.8386	0.0000
v1657st	97000	0.0024	-0.1496	1.4185	-5.4797	6.3623	0.5033	3.2397	
v1659st	97000	-0.0020	0.0126	1.4125	-5.6067	5.5050	-0.0480	2.9513	0.0000
v1698st	97000	-0.0021	0.1632	1.4133	-6.6262	4.5798	-0.7805	4.0702	
v1700st	97000	-0.0007	-0.0312	1.4116	-6.8147	6.0794	-0.0105	3.2498	0.0000
mv1748st	97000	0.0012	0.0898	1.4091	-6.9065	4.9561	-0.3314	3.1100	
v1752st	97000	-0.0028	-0.0242	1.4138	-5.9970	6.0278	0.0584	2.8602	0.0000
v1757st	97000	-0.0016	0.0619	1.4093	-5.6093	6.0796	-0.2061	2.8815	0.0000
mv1759st	97000	-0.0019	0.0019	1.4095	-5.7851	8.4467	0.3504	4.5707	
v1761st	97000	0.0046	-0.0841	1.4092	-4.9151	6.9846	0.3560	3.1644	
v1765st	97000	-0.0044	0.0623	1.4108	-6.3622	5.3760	-0.2583	2.9938	0.0000
v1715st	97000	-0.0004	-0.0465	1.4125	-5.6903	6.4680	0.1635	3.0390	0.0000
v1716st	97000	0.0048	-0.0767	1.4058	-5.4289	8.6264	0.6894	4.8246	
v1719st	97000	0.0028	0.0047	1.4119	-5.8482	5.8327	-0.0015	2.9541	0.0119
v1665st	97000	-0.0021	-0.1498	1.4121	-4.5091	8.1223	1.1296	5.8835	
v1679st	97000	0.0001	-0.1331	1.4067	-5.0659	8.7945	0.8529	4.9128	
v1689st	97000	-0.0019	0.1172	1.4099	-6.3775	5.6745	-0.3794	3.0707	
v1691st	97000	-0.0022	0.0299	1.4072	-5.6079	5.6826	-0.0899	2.9000	0.0000

v1694st	97000	0.0036	0.1461	1.4079	-9.4557	4.9859	-1.3322	7.6375	
mv1769st	97000	0.0035	-0.0367	1.4096	-5.6920	6.7935	0.1479	3.1397	0.0000
mv1772st	97000	0.0030	0.0788	1.4112	-7.0022	5.6204	-0.3424	3.3253	
v1774st	97000	0.0002	-0.0457	1.4080	-5.9041	7.3899	0.2491	3.3162	0.0000
v1777st	97000	0.0010	0.1148	1.4062	-8.7976	5.1049	-0.7608	4.6563	
v1778st	97000	0.0012	0.0192	1.4104	-5.7255	6.0675	-0.0630	2.8926	0.0000
v1725st	97000	0.0010	-0.0873	1.4063	-5.5752	7.3183	0.3759	3.3321	
mv1733st	97000	0.0007	0.0978	1.4076	-6.5336	5.2535	-0.3850	3.3000	
v1740st	97000	-0.0035	0.0822	1.4098	-6.3034	4.8329	-0.3450	3.1448	
v1742st	97000	0.0023	-0.1041	1.4068	-5.5348	7.4721	0.5603	3.8374	
Composite	97000	0.0075	-0.0136	8.3278	-38.5602	34.8713	-0.0697	3.1500	0.0000

The following two figures show the ranking of the 117 hospitals and 97 local authorities on the new composite shown in order from the worst to the best performer (the dark dots). If the simulated data were not produced one might conclude that the performance of the best organisation on the composite appears to be significantly better than the performance of the worst organisation because the dark dots assume all variation is due to differences in performance (no random variation).

Around each of the dark dots the vertical line shows the 95% confidence interval arising from the simulations (the range in which the composite for this organisation could potentially fall 95 percent of the time) – this assumes all variation is random (none due to true performance variation). 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. One cannot be certain that (say) the hospital or local authority ranked 10 is necessarily performing better on the composite constructed for (say) the hospital or local authority ranked 50.

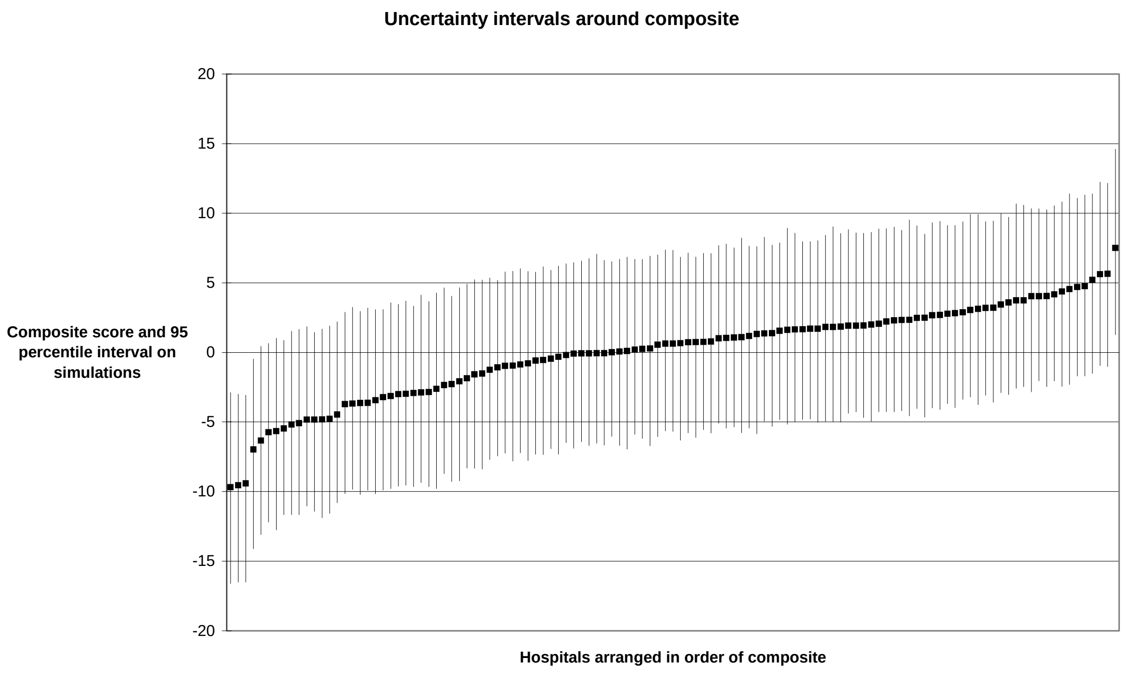


Figure 2: Uncertainty intervals around composite score using simulations, healthcare

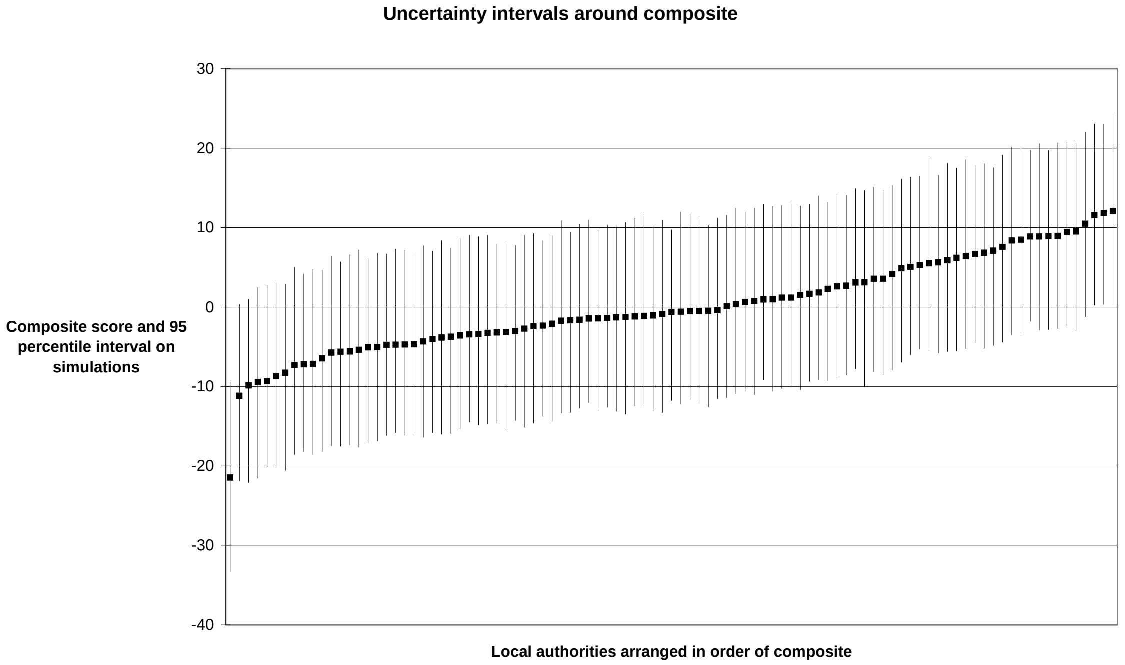


Figure 3: Uncertainty intervals around composite score using simulations, local government

Similarly, when examining the rankings rather than scores of organisations, we obtain the results in the following two figures. Hospitals and local authorities are ranked from worst (117th for hospitals and 97th for local authorities) to best (1st) and the vertical lines show the 95% confidence intervals around these rankings. These illustrate the high degree of uncertainty in the rankings when we assume all variation is random with almost all confidence intervals overlapping.

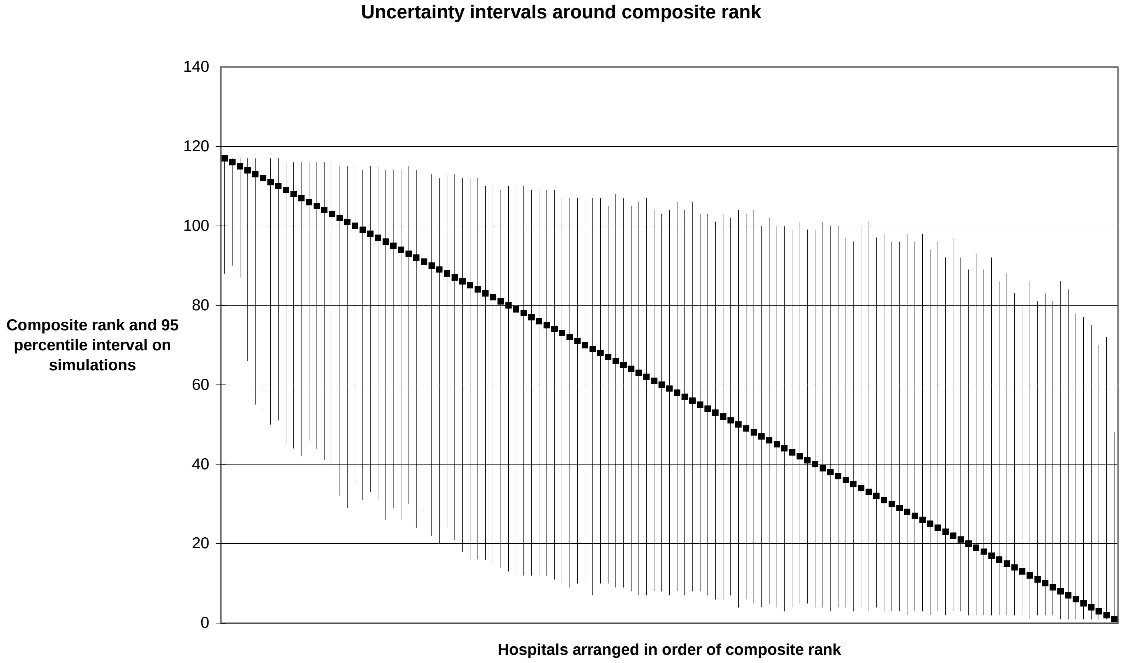


Figure 4: Uncertainty intervals around composite rank using simulations, healthcare

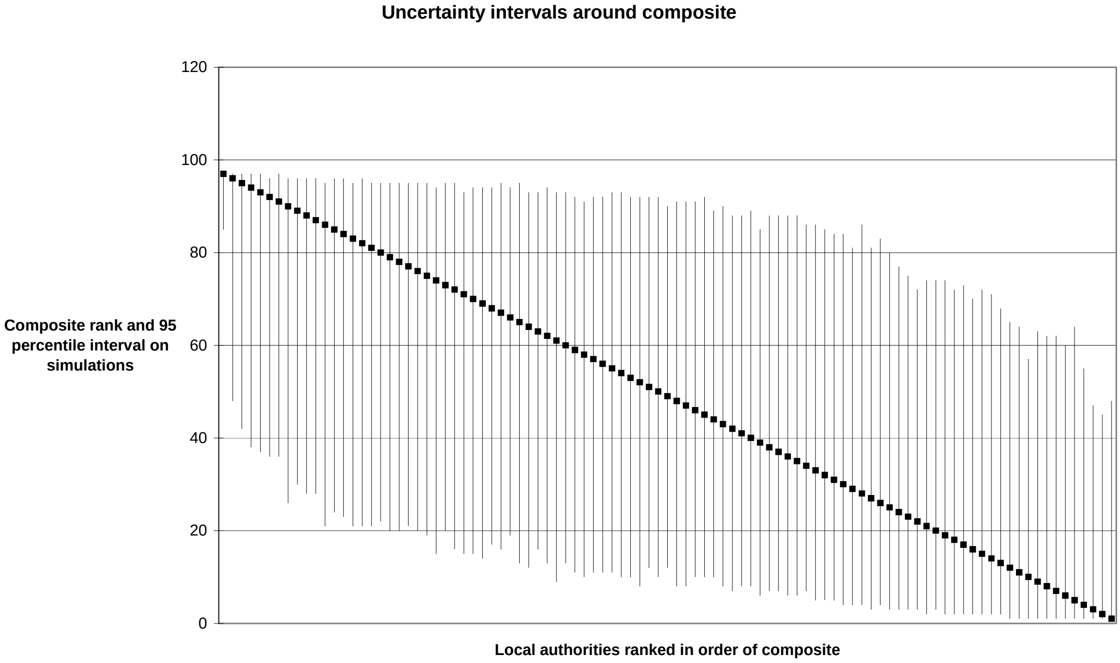


Figure 5: Uncertainty intervals around composite rank using simulations, local government

4.2.4 Random variation on performance indicators

Some of the variation on measured performance will be random variation, the result of measurement error or sampling error or simply natural variation that exists within any distribution. It is important to be able to disentangle different sources of variation on performance measures. For each of the performance indicators, we sought to estimate the proportion of variation caused by factors such as measurement error and random fluctuations. If it were possible to decompose the variation into these elements, as opposed to the true performance variation, one may be able to say with greater certainty what the true differences in performance are between organisations.

While performance on the underlying indicators may be identifiable to some extent, much of the variation in performance across organisations on any particular indicator will be due to a number of indecipherable and unpredictable events such as measurement error or sampling error or simply normal random variation. When comparisons are therefore made between hospitals or local authorities and over time, account must be taken of the variability in performance due to such random events. In practice this means one must know something about how a purely random process generates a distribution of events. When considering different types of indicator variables, many of them produce counts of events (such as deaths or readmissions) and are assumed to emanate from a Poisson distribution which assumes that the events are only subject to random variation. However, it is highly unlikely that any of these indicators will have random variation as the only source of variation. Being able to disentangle these different sources of variation in the indicators (which will translate through into the construction of the composite) will make the use of the uncertainty intervals much more useful since much tighter intervals can be obtained if some of the variation can be taken out. The objective is therefore to obtain an estimate of the variation on each indicator which is due to random variation (measurement error and natural variation). The remainder is assumed to represent true variation in performance.

There is no agreed methodology to do this and as such our proposed methodology seeks to exploit the *within* and *between* variation that exists when longitudinal data is available. While we have panel data for both sectors, the changes in the inclusion and exclusion of certain variables, means that for some indicators longitudinal data was not available. However, for other indicators, more than three years of data were available. The longer the time series available, the better estimate can be made of the degree of random variation *within* organisations over time and *between* different organisations.

In order to get an estimate of the random variation on each performance indicator, a fixed effects panel data regression was run on each of the variables which have data for more than 1 year. Fixed effects regressions are essentially the same as ordinary least squares regressions when indicator (dummy) variables are included for each of the organisations. The fixed-effects model is:

$$y_{it} = a + x_{it} b + v_i + e_{it} \quad (4)$$

where v_i are the organisation-specific fixed effects and year (time) dummies are included on the right-hand side.

The results for each of these regressions produce an estimate of the within variation explained by the organisation-specific effects and the year dummy variables. This R-squared within variation on each of the indicators was used as an estimate of the proportion of that indicator's variation that is random and so beyond the control of the organisation. The remainder of the variation was assumed to reflect genuine variations in levels of performance.

The within variations for each of the indicators in each of the two sectors are shown in the following tables.

Table 28: Proportion of random variation on underlying performance indicators, healthcare

Variable	Proportion random variation assumed
Percent patients waiting up to 6 months for an inpatient admission	80
Junior doctors working hours	66
Percent patients seen within 13 weeks for outpatient appointment	42
Data quality	34
Inpatient survey satisfaction with coordination of care (average assumed)	32
Deaths within 30 days of emergency surgery	27
Emergency readmission within 28 days following hip fracture	25
Discharge home within 28 days following hip fracture	9
Staff satisfaction survey	2
Sickness absence rate	1

For one of the variables (inpatient satisfaction) for which there was no panel data available, an average figure (from the rest of the estimates) of 32 percent was assumed.

Table 29: Proportion of random variation on underlying performance indicators, local government

Variable	Proportion random variation assumed
Unauthorised absences secondary schools	98
Recycling	60
Types of interactions delivered electronically	58
Pupils 5 or more GCSEs: A*-C	54
Working age (18-65) people with disabilities	36
Composting	31
Perm excluded pupils attending alternatives tuition: 20 hours or more a week	27
Planning major apps in 13 weeks	25
Priv. dwellings 6 months empty - returned to occupation	24
Community Strategy	22
Working age (18-65) people from ethnic minorities	22
Burglaries per 1000 households	19
Cost intensive care adults	18
Housing benefit security - number of investigations per 1000 caseload (average assumed)	18
Care leavers in education / training / employment	15
Il health retirements / staff	14
Sore on creating opportunity checklist	13
Condition classified non-principal roads	12
New homes on brown field sites	8
Needs statements	8
Senior women	7
Standard searches in 10 working days	6
Racial incidents further action	5
Road accidents - car users	4
SENs in 18 weeks with exceptions	4
Footpaths easy to use	3
Principal roads not needing major repair	3
Household waste collected	2
Renewal claims on time	2
Visits to libraries	2
Length of stay in hostels	1
Children with 3+ placements	1
Schools subject to special measures	1
Secondary schools 25% + surplus places	1
Older people helped to live at home	1

For one of the variables (housing benefit security) for which there was no panel data available, an average figure (from the rest of the estimates) of 18 percent was assumed.

A similar pattern is found in the degree of random variation in the underlying individual indicators when comparing results across the two sectors. For healthcare it varies from 80% to 1%. For local authorities it ranges from 98% to 1%. There is a similarly wide range in the estimated proportion of variation across the 10 hospital performance indicators and the 35 local authority performance indicators. Some of this variation may be driven by slight changes in definition to the indicators over time, subtle changes in the way the data is collected or measured over time, performance targets attached to individual indicators which may lead to increased “within” variation as organisations improve their performance over time, and possible “gaming” behaviour. Differences in the degree of variation may also be the result of certain indicators being more or less subject to managerial control and mediation and less subject to random events (Hauck *et al*, 2003).

To show the impact of random variation on the individual indicators, we used the *within* variation estimates to “shrink” the standard deviation for each indicator so that the mean of the distribution is its reported performance in 2001/02 (healthcare) and 2003/04 (local government) and its standard deviation is the estimated proportion of random variation which remains on that indicator as listed in the tables above. In the example in the following figure, the variation around the distribution for survival rates reduces from ~ N(0,1) (which assumes all variation is random) to ~ N(0,0.27), where 27 percent is the estimate of variation that is random on survival rates in hospitals. This was done for all 10 performance indicators for healthcare and all 35 performance indicators for local government.

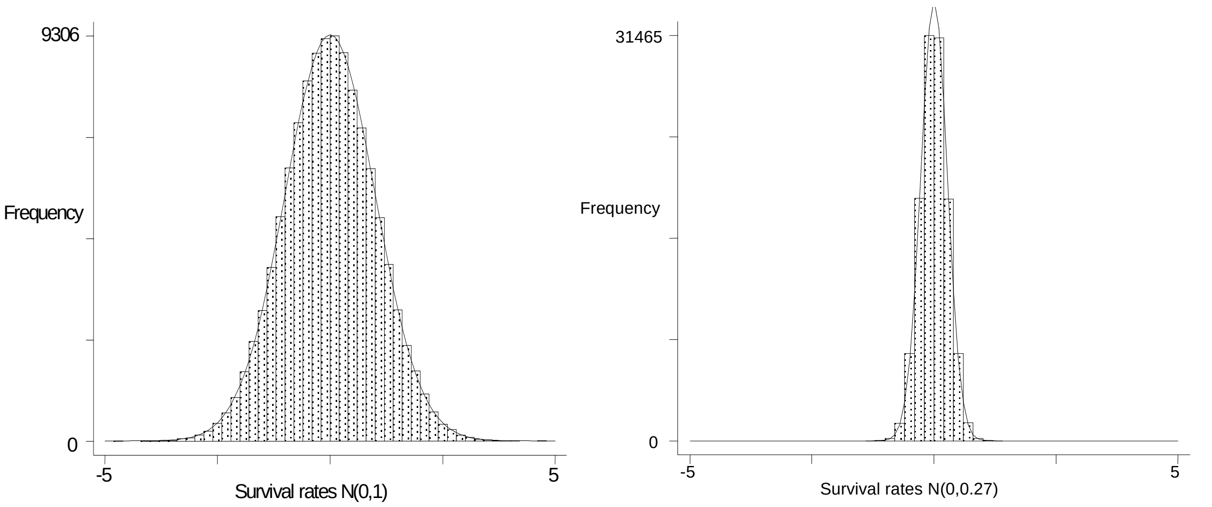


Figure 6: Example of variance around survival rates with 1000 observations with distribution N(0,1) and N(0,0.27)

The variables with the reduced variation were now used to re-construct the uncertainty intervals around the composite indicator. While the black dots of the original composite remain unchanged in the following figures, the variation around the composite has now shrunk dramatically as a result of the reduction in variation around the underlying indicators to the proportion of variation due to random events.

After taking account of random variation in the underlying performance indicators, the results illustrate that we were able to estimate genuine performance variations. It is now possible to say with certainty that (for example) the organisations in the bottom quartile are scoring less well than those in the top quartile, though there is still overlap in the middle.

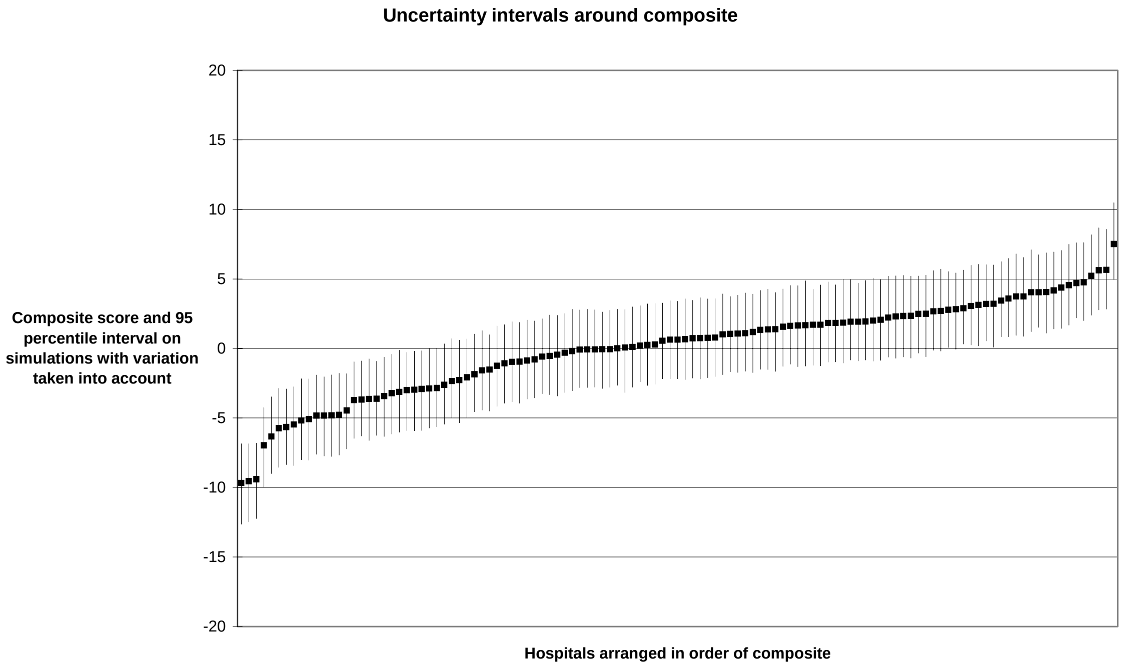


Figure 7: Uncertainty intervals around composite score using simulations with random variation taken into account, healthcare

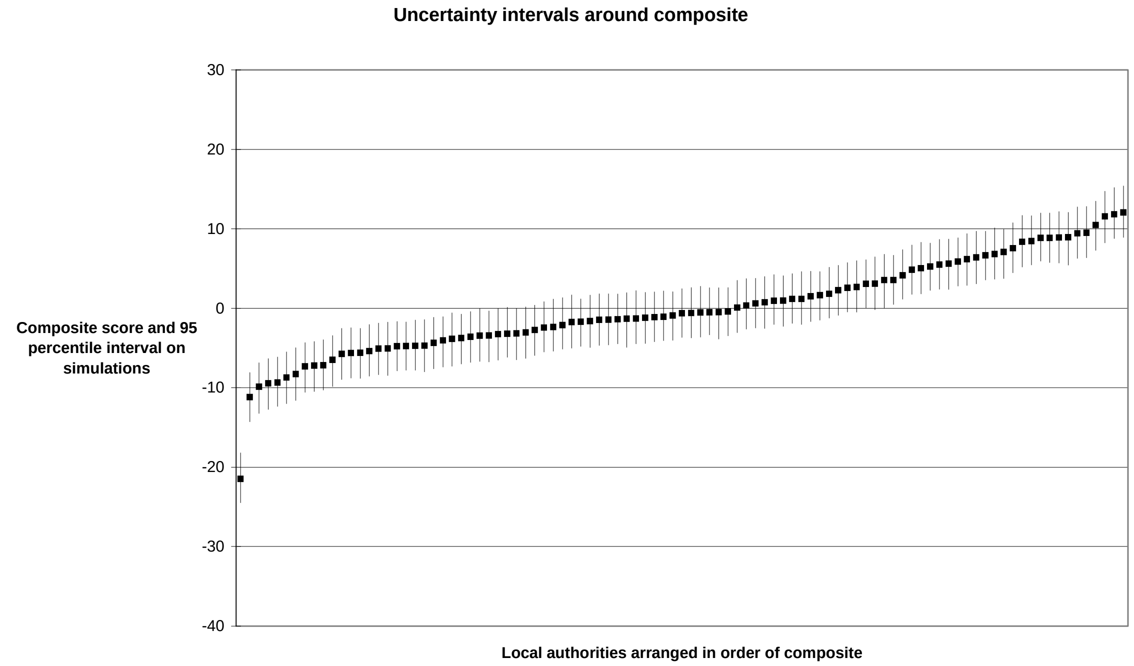


Figure 8: Uncertainty intervals around composite score using simulations with random variation taken into account, local government

The importance of decomposing variation in performance indicators is highlighted as a worthwhile analytical practice since it can dramatically change conclusions about performance.

Similarly, when examining the rankings rather than scores of organisations, we illustrate that the naive view, attributing all variation to randomness, is radically altered (with still some overlap in the middle) after accounting for random variation.

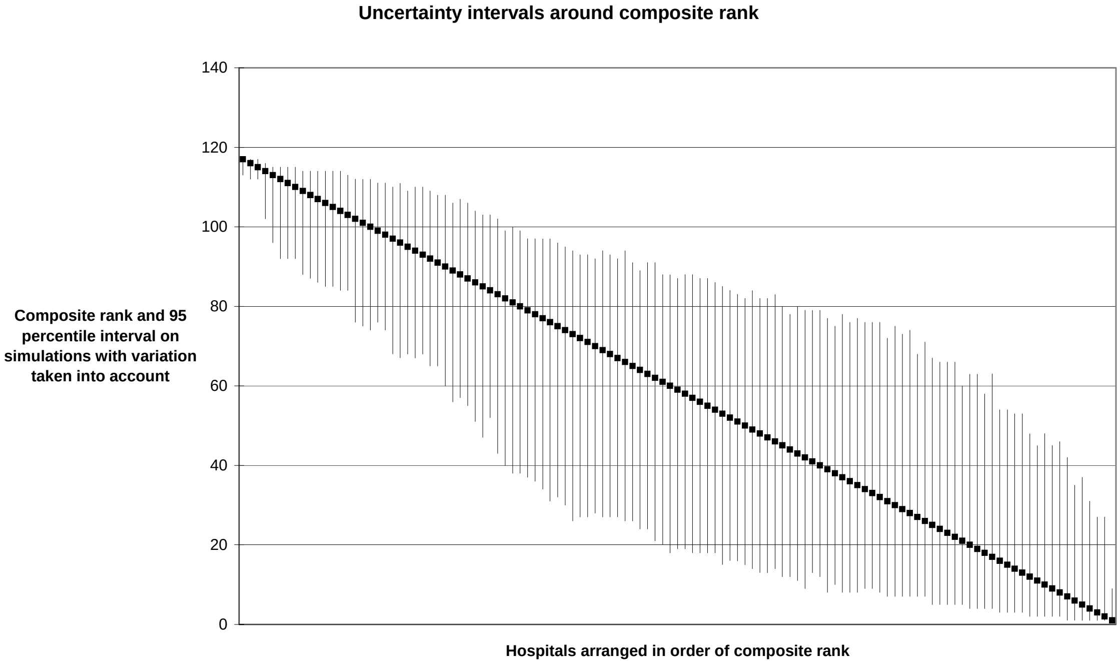


Figure 9: Uncertainty intervals around composite rank using simulations with random variation taken into account, healthcare

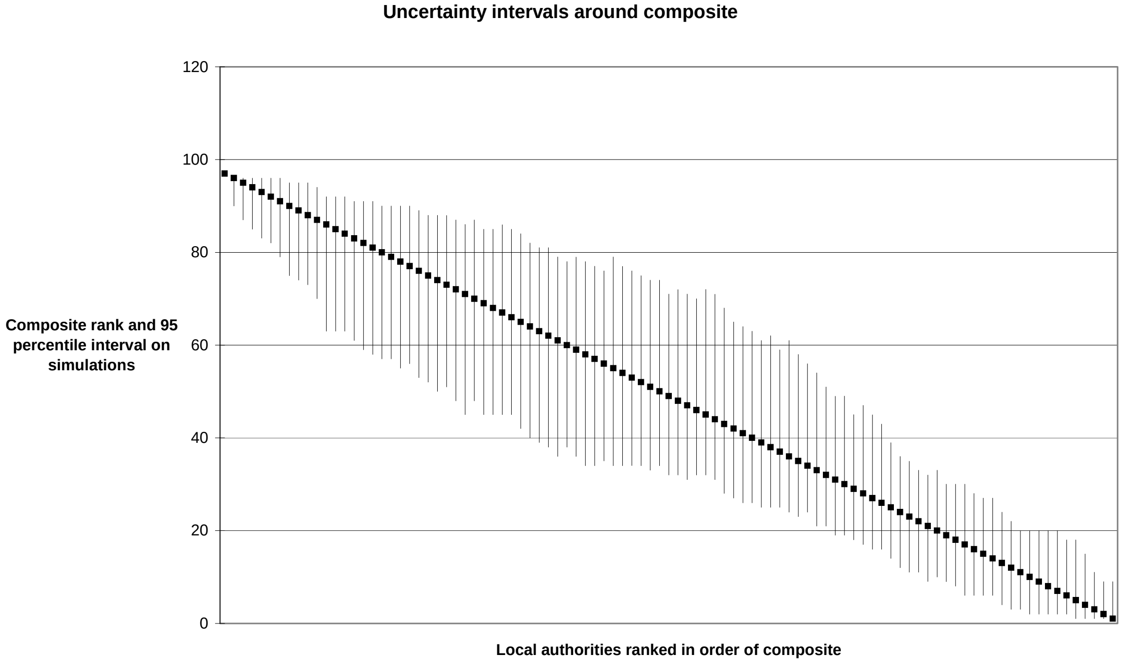


Figure 10: Uncertainty intervals around composite rank using simulations with random variation taken into account, local government

4.2.5 Introducing changes in weights

We explore whether changes in methods to aggregate underlying indicators to construct the composite indicator can have an impact on the results. This section explores the sensitivity of the rankings on the composite index to changes in weighting systems.

We first tested the impact of altering the weights attached to specific indicators, whilst retaining the principle of linear aggregation. As well as affecting composite scores, this also in general affects variability. If we combine *n* random variables (x_j, σ_j^2) with weights *w_j*, the variance of the composite measure is:

$$\sigma^2 = \sum_{j=1}^{n} w_j^2 \sigma_j^2 + \sum_{j>k} \sum_{k=1}^{n} w_j w_k \operatorname{cov}(x_j, x_k) \qquad (5)$$

where *cov* (x_j,x_k) is the covariance between indicators *j* and *k*. The marginal impact of an increase in the weight attached to indicator *j* is therefore given by:

$$\frac{\partial \sigma^2}{\partial w_j} = 2 w_j \sigma_j^2 + 2 \sum_{k \neq i} w_k \operatorname{cov}(x_j, x_k)\,. \qquad (6)$$

So an exogenous increase in weight w_i leads to an *increase* in variance, providing

$$w_j \sigma_j^2 + \sum_{k \neq i} w_k \operatorname{cov}(x_j, x_k) > 0\,. \qquad (7)$$

Usually this will be the case, as 'good' organizations tend to score well across most indicators (that is, cov(x_j,x_k)>0 for most (*j*,*k*)).

We are interested only in increases in the *relative* weight attached to indicator *i*. This can be readily accommodated by dividing each weight by a scaling factor:

$$W = \sum_{i=1}^{n} w_i \qquad (8)$$

Then:

$$\sigma_C^2 = \left\{ \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} w_i w_j \operatorname{cov}(x_i, x_j) \right\} / W^2 \qquad (9)$$

and

$$\frac{\partial \sigma_C^2}{\partial w_i} = \frac{2}{W^2} \left\{ w_i \sigma_i^2 + \sum_{j \neq i} w_j \operatorname{cov}(x_i, x_j) \right\} \left(1 - \frac{w_i}{W} \right), \qquad (10)$$

ensuring that the weights are normalised and always add to the same total.

The composite index with random variation taken into account (thus showing true performance variation), is used to explore changes in the weights applied to the underlying indicators. When the weight of a particular variable is changed, the standard deviation of that particular variable changes according to the new weight, thus if the weight is (say) doubled (×2), the standard deviation also doubles (1×2), assuming a z score (0,1) distribution. The composite still has zero mean but the standard deviation (and variance) also increases, although it has a larger standard deviation than the underlying indicator. Of course the opposite happens if the weight of a particular variable is reduced, its standard deviation reduces by the chosen amount, and the variance on the composite also decreases.

A large number of scenarios were tested for changes in weighting structures, in particular exploring the relationship between the correlation of variables and changes in the weights applied to them.

By way of illustration, we summarise in the following table for healthcare, the impact of changing the weights on:

1) the three outcome indicators (death rates following emergency surgery, readmission rates for hip fracture, discharge home following hip fracture) and the two access measures (thirteen week outpatient waits, and six month inpatient waits), and
2) the two access measures

These are increased and decreased by a factor of 2 and 3 respectively. We show the impact on the rankings for the top, middle and bottom 5 hospitals from the original rankings. This confirms that the very good organisations are not likely to be confused with the very poor ones. But middle-ranking hospitals are poorly differentiated and unstably ranked. The correlation between the new and original rankings varies between 0.88 and 0.96. The largest jump in position for an individual hospital was 54 places, almost half the league table. On average, hospitals changed between 7 and 13 places in the rankings, depending on the changes made to the weighting system.

Table 30: Impact of different weighting systems on hospitals' composite rankings (with 95% uncertainty intervals) for top, middle and bottom 5 ranked hospitals

Hospitals	Original composite index (equal weights)			5 weights halved			5 weights doubled			2 weights reduced by two-thirds			2 weights tripled		
	Rank	95% CI		Rank	95% CI		Rank	95% CI		Rank	95% CI		Rank	95% CI	
Top 5	1	1	1	1	1	5	1	1	1	2	1	4	2	1	3
	2	2	8	4	1	8	4	2	12	6	3	15	4	3	9
	3	2	8	2	1	7	6	2	12	1	1	4	18	10	34
	4	2	10	6	2	11	5	2	12	8	3	15	9	4	13
	5	2	14	5	2	11	10	4	20	14	6	23	8	4	11
Middle 5	57	38	74	68	57	83	40	22	61	55	35	72	57	39	73
	58	39	75	42	28	61	69	50	81	31	21	56	80	66	92
	59	39	74	51	35	68	65	42	75	34	23	56	78	62	90
	60	39	75	65	45	76	54	30	70	42	25	62	72	56	85
	61	41	75	63	42	72	63	39	73	79	63	87	26	14	45
Bottom 5	113	108	114	112	109	113	112	104	114	110	106	113	113	109	116
	114	112	114	115	114	117	105	95	112	114	113	115	106	97	111
	115	115	117	114	114	117	117	115	117	116	116	117	114	110	116
	116	115	117	116	114	117	115	115	117	117	116	117	111	101	113
	117	115	117	117	114	117	116	115	117	115	114	115	117	117	117
Correlation between original and new rankings	-			0.96			0.96			0.93			0.88		
Largest change in ranking	-			39 places			34 places			42 places			54 places		
Average change in ranking	-			7 places			7 places			9 places			13 places		

As suggested by these examples, the weighting structure can indeed materially affect the rankings of hospitals on the composite. Predictably, the ranks change the most for units when weights for indicators on which they perform exceptionally (well or poorly) are increased or decreased. Changing the weight structure and the impact this ultimately has on rankings, is related to the degree of correlation between the underlying indicators. This relates to the formal relationship given in equations (9) and (10) which show that the change in weights is intimately linked to the covariance between indicators.

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

Table 31: Weighting applied in the Comprehensive Performance Assessment (CPA)

Seven domains:	Weight:
Education	4
Social services	4
Environment	2
Housing	2
Libraries and leisure	1
Benefits	1
Use of resources	1

We explored the impact on the original composite indicator of changing the weights on the underlying performance indicators. The following table 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 the above table for CPA. The results highlight the change in ranking across 97 places for the top, middle and bottom 5 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 was 54 places, more than half the league table. On average, authorities changed 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.

Table 32: Impact of different weighting systems on local authorities' composite rankings for top, middle and bottom 5 ranked authorities

	Original	Ed+SS x4	Ed+SS x0.25	Env+Ho x2	Env+Ho x0.5	Ed+SS x4 + Env+Ho x2 + Rest x1
Top 5	1	3	4	1	9	9
	2	21	1	2	5	10
	3	2	13	11	1	4
	4	5	7	5	4	3
	5	30	3	8	7	33
Middle 5	47	42	62	35	57	35
	48	58	43	51	49	75
	49	55	52	30	64	60
	50	54	54	62	47	55
	51	86	28	31	67	70
Bottom 5	93	84	91	90	95	90
	94	79	93	91	94	93
	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

4.2.6 Introducing decision rules

In this section we explore the impact on uncertainty of applying a set of sequential non-linear decision rules to aggregate the indicators into ordinal categories of performance. Decision rules are used 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. Such decision rules are frequently used by policymakers when they wish (for example) to penalize failure to achieve minimum standards. We use a set of rules which emulates the process used to construct the Star ratings and CPA in as much as hospitals and local authorities have to meet a minimum threshold of performance on certain indicators in order to achieve higher Star or CPA ratings. A number of different combinations and permutations of decision rules were tested.

Prior to applying the decision rules, the variables were first transformed into categorical variables similar to those applied in the Star and CPA ratings, with a scale of 1 to 3 (for healthcare) and 1 to 4 (for local government), reflecting the desire that certain thresholds of performance be achieved on each indicator. For example, in the Star ratings system, hospitals are categorized according to whether they have ‘achieved’, ‘underachieved’ or ‘significantly underachieved’ standards for each of the key targets. Similarly, in the CPA, services are given a score of 1 to 4. In this example the same thresholds are applied to all 10 indicators for healthcare and all 35 indicators for local government. Each of the standardised variables was transformed into a categorical variable representing thresholds of performance.

There are 3 categories for healthcare, those with a standardised score above 0.5 were assigned a 3, those with a score below –0.5 a 1 and the remainder a score of 2. This partitions the indicators into three levels of performance with roughly equal numbers of hospitals in each.

There are 4 categories for local government, those with a standardised score above 0.7 were assigned a 4, those between 0 and 0.7 a 3, those between 0 and –0.7 a score of 2, and the remainder with a score below –0.7 a 1. This partitions the indicators into four levels of performance with roughly equal numbers of local authorities in each.

The decision rules that are often employed in the construction of composites were then simulated. Thus a new composite index was constructed, like the Star ratings and CPA, with four categories from zero to three (for healthcare) and five categories from zero to four (for local government) based on the above categorical variable transformations. The rules were chosen based on the fact that the variables were transformed to categorical variables. A number of different combinations and permutations of decision rules were tried.

An example of the type of decision rules which were sequentially applied to the transformed data is given in the following two tables for healthcare and local government respectively.

Table 33: Sequential decision rules applied to create the illustrative composite measure, healthcare

Performance on the 10 variables was first transformed into categorical variables (as for star ratings) on a scale of 1 to 3.

THEN:

3 stars IF achieve a score of 2 or 3 on ALL three of the outcome variables: readmission rate following hip fracture; death rates following surgery; discharge home following hip fracture AND on the two waiting times variables: 13 week outpatient waits; six month inpatient waits

0 stars IF achieve a score of 1 or 2 on ALL of these three staff and patient variables: inpatient satisfaction; staff satisfaction; junior doctor hours

1 star for all other hospitals UNLESS they achieve a score of 3 on EITHER of: data quality; sickness absence, in which case they get 2 stars.

Table 34: Sequential decision rules applied to create the illustrative composite measure, local government

Performance on the 35 variables was first transformed into categorical variables (as for star ratings) on a scale of 1 to 4.

THEN:

4 stars IF achieve a score of 3 or 4 on ALL education variables OR achieve a score of 3 or 4 on ALL social services variables OR achieve a score of 4 on ALL environment variables OR achieve a score of 4 on ALL housing variables OR achieve a score of 4 on ALL planning variables OR achieve a score of 4 on ALL transport variables.

3 stars IF achieve a score of 3 or 4 on ALL corporate health variables OR achieve a score of 3 or 4 on ALL benefits variables.

1 star IF achieve a score of 1 or 2 on ALL environment variables OR achieve a score of 1 or 2 on ALL housing variables OR achieve a score of 1 or 2 on ALL planning variables OR achieve a score of 1 or 2 on ALL transport variables.

0 star IF achieve a score of 1 on ALL community safety variables.

2 star for all other local authorities.

The composite was constructed in the same way as before, as a simple linear summation of the underlying indicators, now in categorical form.

Applying the rules of the algorithm in the order above effectively places a set of implicit weights on the variables which are used to dictate the thresholds for best and worst performance. Thus certain variables will implicitly receive a higher weighting, which will therefore impact on the rankings of organisations.

In the different scenarios that were tested, it is clear that subtle and highly subjective changes to the decision rules can dramatically impact on how many organisations end up in each category. These decision rules were chosen so as to try to ensure an approximately equal number of organisations in each group. However, the analyst could easily change the number of organisations in each category by simply changing the rules in subtle ways from (say): Organisations receive a three star if they achieve a three on *all* of the three performance indicators to (say): Organisations receive a three star if they achieve a three on *any* of the three performance indicators.

The following figures show the new (categorical) composites based on the above decision rules for healthcare and local government respectively. The new composite (black dots) therefore take values of exactly 0, 1, 2 or 3 only (for healthcare) and 0, 1, 2, 3 or 4 only (for local government) and the uncertainty intervals will equally cover (potentially) the same range. The five categories for local government represent 0 to 4 stars in 2004/05 and the five categories “excellent, good, fair, weak, poor” in the years 2001/02 – 2003/04. These are assumed to be equivalent, such that “excellent” equals 4 stars, “good” equals 3 stars, and so on.

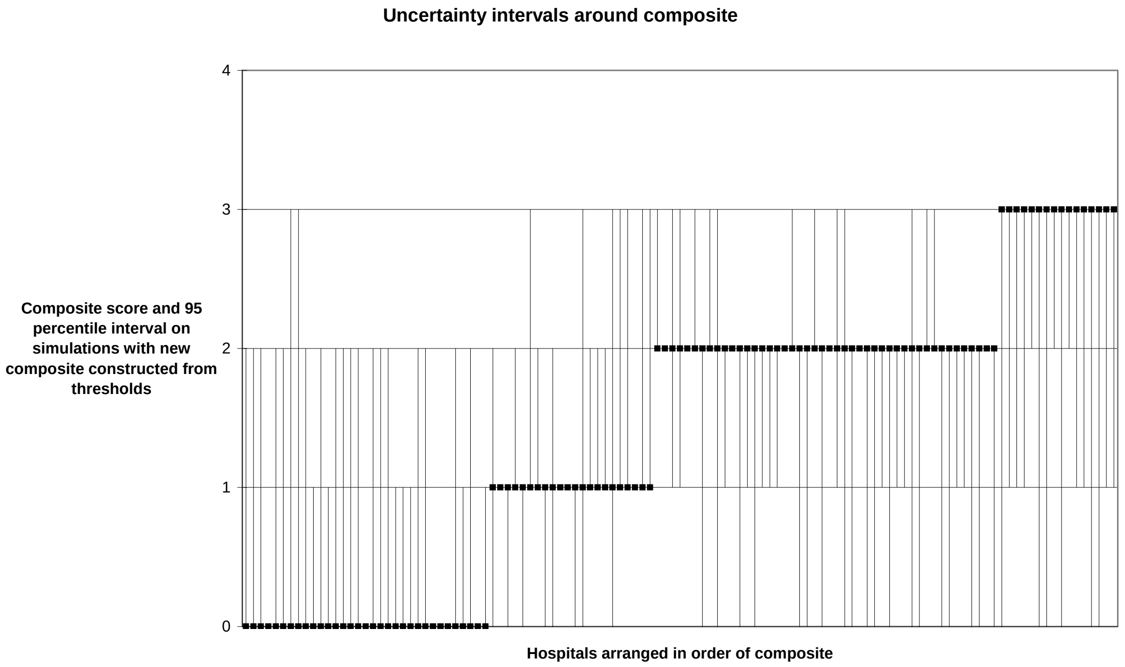


Figure 11: Uncertainty intervals around a new composite score constructed from thresholds, healthcare

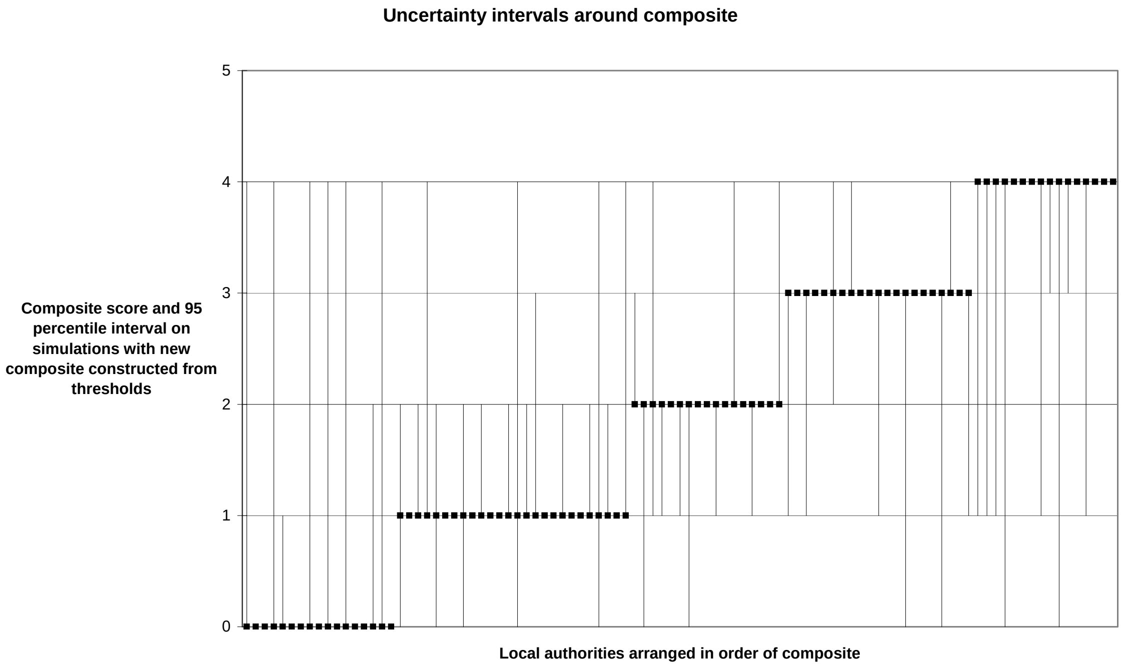


Figure 12: Uncertainty intervals around a new composite score constructed from thresholds, local government

The potential for units to change ranking is dramatic when these sorts of decision rules or algorithms are applied to construct a composite index and potentially small and subtle changes to the rules can materially affect the outcome for individual organisations.

In the next example, rather than generate conventional confidence intervals, the Monte Carlo simulations were used to estimate the probability that an organisation would be found in each of the performance categories.

The following table shows the frequency distribution for the number of times that the hospitals are ranked in each of the categories. 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, randomly selected from each of the 4 categories (based on the original assignment to star category). We found that whilst there was relative stability for the worst performing hospitals (with zero score), this was not the case for the best performers (three stars). So for example, hospitals 1 and 2 achieved a zero score in 100% and 82% of the simulations respectively; whereas hospitals 9 and 10 received the score of 3 stars only 61% and 56% of the time respectively. Indeed, because of the threshold nature of the decision rules, hospital 9 suffers a catastrophic relegation to a zero score in 38% of the simulations.

Table 35: Frequency distribution of a sample of 10 hospitals on the new composite index constructed from decision rules

Hospital	New composite category	Percentage of times in simulations that composite is given a score of:			
		0	1	2	3
1	0	**100**	0	0	0
2	0	**82**	18	0	0
3	0	**66**	0	34	0
4	1	2	**61**	0	38
5	1	0	**100**	0	0
6	2	38	2	**32**	28
7	2	19	0	**81**	0
8	2	0	0	**100**	0
9	3	38	2	0	**61**
10	3	0	0	44	**56**

The following table illustrates the frequency distribution for the number of times a local authority is placed in each of the five categories, using a sample of 15 local authorities, randomly selected from each of the 5 categories “excellent, good, fair, weak, poor” (based on the original assignment to category).

While the picture is not as clear cut in terms of relative stability for the worst performing local authorities, there are still potentially large changes in categorisation possible, for example authority 13 receives a rating of “excellent” for 57% of the time, but a rating of “weak” for 41% of the time.

Table 36: Frequency distribution of a sample of 15 local authorities on the new composite index constructed from decision rules

Local authority	New composite category	Percentage of times in simulations that composite is given a score of:				
		Poor	Weak	Fair	Good	Excellent
1	Poor	**93**	0	0	0	6
2	Poor	**100**	0	0	0	0
3	Poor	**51**	16	0	30	3
4	Weak	13	**84**	3	0	0
5	Weak	0	**33**	5	33	30
6	Weak	3	**79**	11	0	8
7	Fair	28	0	**72**	0	0
8	Fair	0	2	**94**	0	3
9	Fair	0	0	**68**	24	8
10	Good	0	2	29	**39**	30
11	Good	0	0	0	**100**	0
12	Good	0	0	0	**68**	32
13	Excellent	0	41	2	0	**57**
14	Excellent	3	0	0	0	**97**
15	Excellent	0	0	0	50	**51**

The following two tables summarise the results for all hospitals and local authorities. The first highlights very clearly for healthcare the greater stability in the ranking of the worst hospitals (0 stars) over the 1000 simulations – 79% of the time they remain a 0 star, whereas this drops to 54% of the time where they remain a 3 star.

While there is a similar gradient in local authorities, it is less pronounced, dropping from 88% for “poor” ratings to 85% for “excellent” ratings.

Table 37: Frequency distribution of hospitals on the new composite index constructed from decision rules

Number of hospitals	New composite category	Percentage of times in simulations that composite 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**

Table 38: Frequency distribution of local authorities on the new composite index constructed from decision rules

Number of local authorities	New composite category	Percentage of times in simulations that composite is given a score of:				
		Poor	Weak	Fair	Good	Excellent
17	Poor	**87.6**	3.6	0.2	3.7	4.9
26	Weak	1.4	**85.7**	8.4	1.6	2.8
17	Fair	3.7	6.3	**85.9**	2.7	1.3
21	Good	2.6	6.8	1.8	**84.5**	4.3
16	Excellent	0.4	6.3	1.2	7.4	**84.7**

The above results suggest that the stability of the composite constructed from decision rules is not always very high.

4.2.7 Changes in ratings 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 alone, rather than methodology. Thus in order to explore the stability of ratings over time, we replicated the process of generating a generic composite indicator for each sector by repeating the analyses for each year of available data in each sector (2000/01 to 2004/05 for healthcare and 2001/02 to 2004/05 for local government), assuming the methodology to construct the composites is constant over time.

Since the choice of indicators included in each composite (the Star ratings and CPA) has changed over time, we had to assume some indicators in some periods time-invariant. The following two tables show the data availability (marked with an X) for each chosen indicator over each of the years. The missing data was thus imputed as being the same as the adjacent year.

Table 39: Data availability over time for the 10 chosen indicators, healthcare

Variable	2000/01	2001/02	2002/03	2003/04	2004/05
Percent patients waiting up to 6 months for an inpatient admission		X	X	X	X
Junior doctors working hours		X	X	X	X
Percent patients seen within 13 weeks for outpatient appointment	X	X	X	X	X
Data quality	X	X	X		X
Inpatient survey satisfaction with coordination of care		X			
Deaths within 30 days of emergency surgery	X	X		X	
Emergency readmission within 28 days following hip fracture	X	X	X	X	X
Discharge home within 28 days following hip fracture	X	X			
Staff satisfaction survey		X	X		
Sickness absence rate	X	X	X		X

Table 40: Data availability over time for the 35 chosen indicators, local government

Variable	2001/02	2002/03	2003/04	2004/05
Renewal claims on time BV78c	X	X	X	
Housing Benefit Security BV76c			X	X
Burglaries per 1000 households BV126a		X	X	X
Racial incidents further action BV175	X	X	X	X
Community Strategy BV1a		X	X	
Senior women BV11a		X	X	X
Ill health retirements / staff BV15	X	X	X	X
Working age (18-65) people with disabilities BV16b	X	X	X	X
Working age (18-65) people from ethnic minorities BV17b	X	X	X	X
Types of interactions delivered electronically BV157	X	X	X	X
Score on Creating Opportunities checklist BV114		X	X	
Visits to libraries BV117	X	X	X	X
Secondary schools 25% + surplus places BV34b	X	X	X	X
Pupils 5 or more GCSEs: A*-C BV38	X	X	X	X
SENs in 18 weeks with exceptions BV43b	X	X	X	X
Unauthorised absences secondary schools BV45	X	X	X	X
Schools subject to special measures BV48	X	X	X	X
Perm. excluded pupils alternative tuition: 20 hours BV159d		X	X	X
Recycling BV82a	X	X	X	X
Composting BV82b	X	X	X	X
Household waste collected BV84	X	X	X	X
Priv. dwellings 6 months empty: returned to occupation BV64	X	X	X	X
Length of stay in hostels BV183b		X	X	X
New homes on brown field sites BV106	X	X	X	X
Major planning apps in 13 weeks BV109a		X	X	X
Standard searches 10 working days BV179	X	X	X	X
Children with 3 + placements BV49	X	X	X	X
Cost intensive care for adults BV52	X	X	X	X
Older people helped to live at home BV54	X	X	X	X
Needs statements BV58	X	X	X	X
Care leavers in education / training / employment BV161	X	X	X	X
Condition classified non-principal roads BV97a	X	X	X	X
Road accidents: car users BV99d	X	X	X	
Footpaths easy to use BV178	X	X	X	X
Principal roads not needing major repair BV186a		X	X	X

The following figures show the change in ratings over time when taking account only of changes in performance data – methodology is assumed constant over time. We compare our results to the actual Star ratings and CPA ratings where methodology (and data) does change year-on-year.

In the actual NHS Star ratings, the methodology (aggregation methods, weighting systems, and decision rules) can change every year. The first figure shows the frequency distribution of 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.

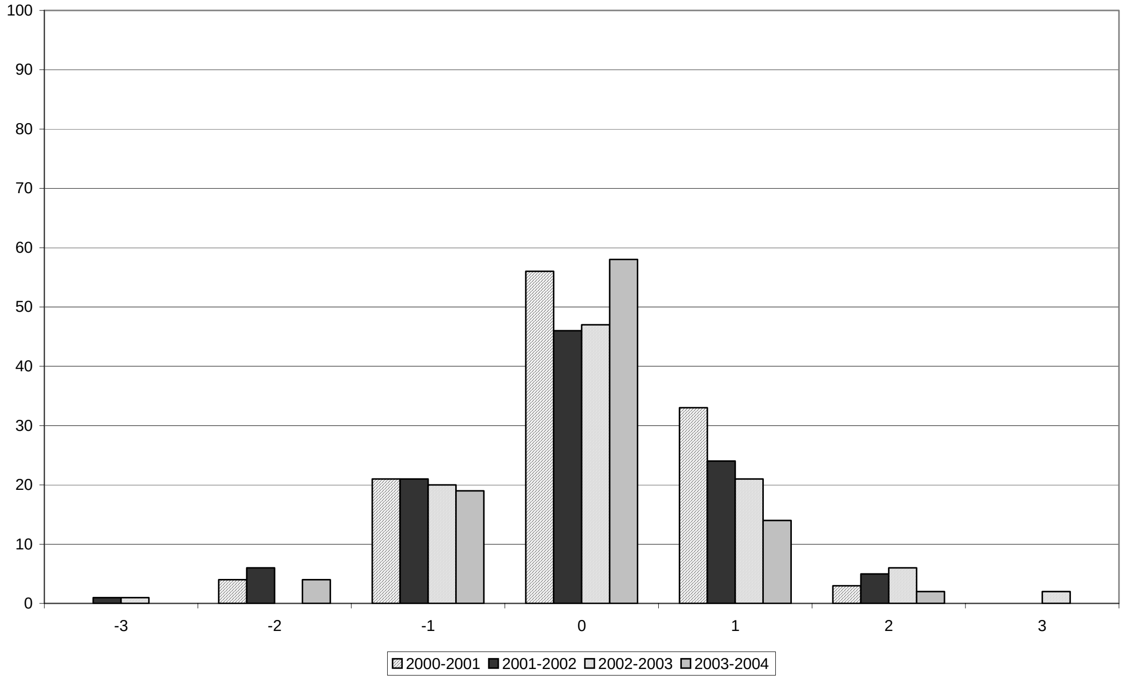


Figure 13: Frequency distribution of hospitals changing ratings on the actual star ratings, n=117

The next figure shows the frequency distribution of changes in our generic composite rating over time for our sample of 117 hospitals.

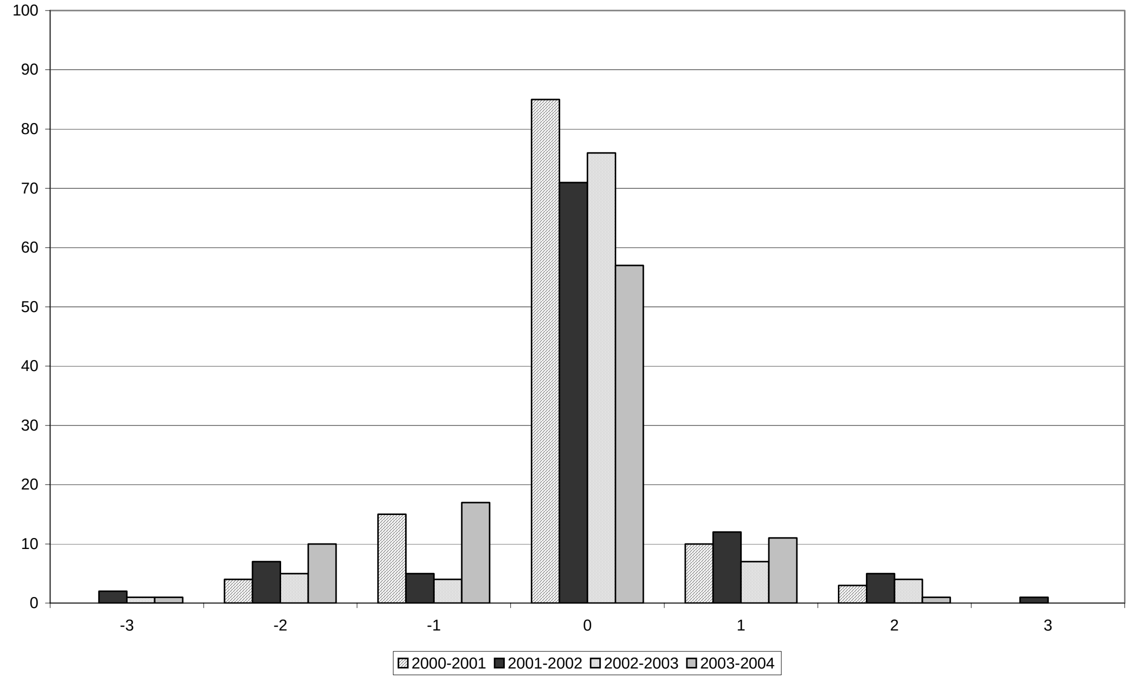


Figure 14: Frequency distribution of hospitals changing ratings on the new composite index constructed from decision rules, n=117

In our composite 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 proportion of changes in ratings is much smaller and is purely due to data fluctuations.

Results were less clear for local authorities. The first figure shows the frequency distribution of changes in the actual CPA ratings over time for our sample of 97 local authorities. The figure illustrates the proportion of authorities in each year which move up or down 1, 2, 3 or 4 categories.

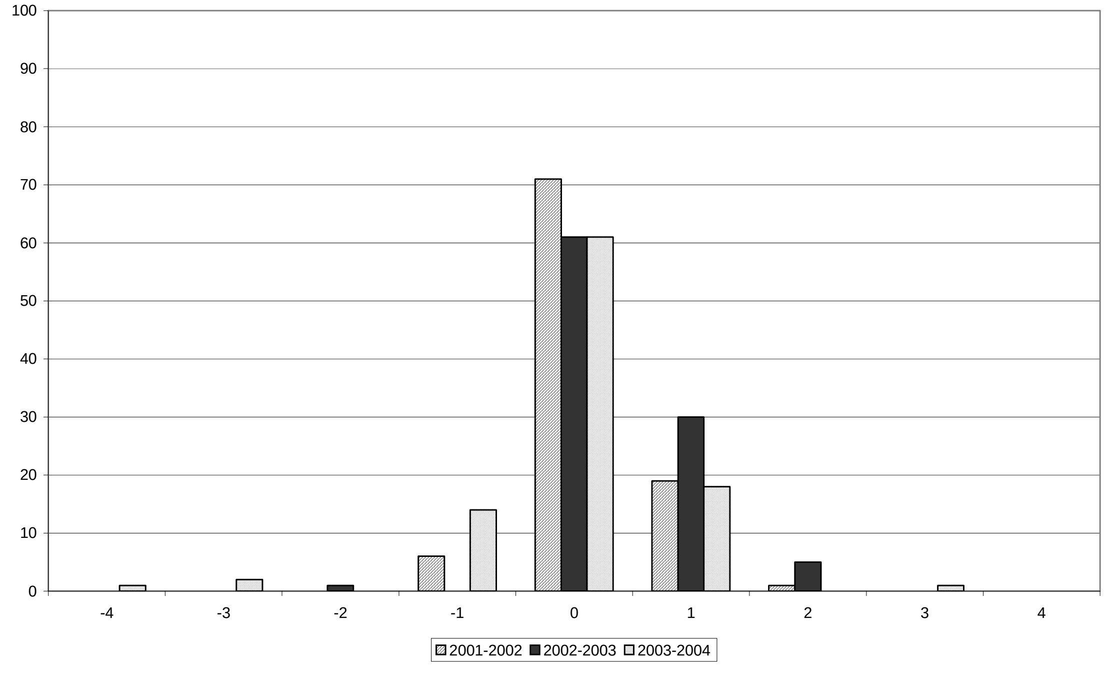


Figure 15: Frequency distribution of local authorities changing ratings on the actual CPA ratings, n=97

The next figure shows the frequency distribution of changes in our generic composite rating over time for our sample of 97 local authorities. The proportion of authorities whose ratings remain unchanged over time (zero) is in fact slightly smaller in our generic composite indicator. This may be due to larger year-on-year data fluctuations in local government.

Nonetheless, a great deal of instability can potentially be introduced into composite ratings over time when there are year-on-year changes to methods, often driven by changes in policy priorities and political imperatives. Organisations may need to learn to adapt to continual changes in the performance measurement system on which they are being rated. Alternatively, continual changes to the system could create difficulties in adopting longer term planning horizons and more strategic approaches to performance improvement. These can have significant consequences for individual organisations if rewards and penalties are attached to the outcomes of the composite indicator.

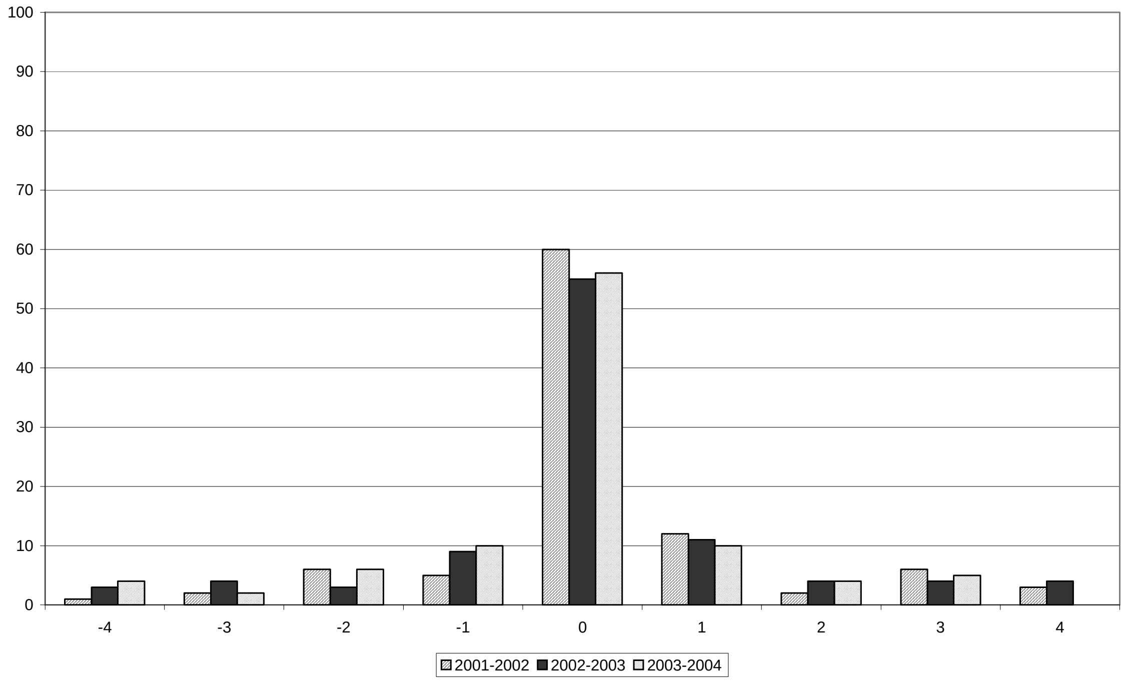


Figure 16: Frequency distribution of local authorities changing ratings on the new composite index constructed from decision rules, n=97

5. Conclusions

Composite indices are a useful communication and political tool for conveying summary performance information in a relatively simple way and signalling policy priorities. They are used widely in various sectors in public services. Composite performance indicators have a number of advantages, such as focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the ‘big picture’ in a way in which the public can understand. It is likely therefore that they will continue to be used in the future in many policy areas.

However, it is important to recognise that the construction of composite indicators is not straightforward and many methodological issues need to be addressed carefully if the results are not to be misinterpreted and manipulated.

These results have important scientific, policy and practice implications.

The key contribution of this research is the disentangling of genuine performance variations from random fluctuation in the measurement of performance indicators and the clear visual illustrations of the impact of uncertainty on performance measures. Our findings will have an important scientific impact since they provide a mechanism whereby researchers and analysts can account for ‘random variation’ in performance indicators more generally. 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 differences in performance. We add to the knowledge base by proposing a useful new methodology to partition the variance and gain precision in performance assessment.

Our results do not however represent merely a methodological or technical pre-occupation. 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 and 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. “Decision rules” need to be treated with caution. Subtle and highly subjective changes to the decision rules can dramatically impact on the composite index and rankings of organisations.

2. 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.

3. 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.

4. The proper treatment of uncertainty in composite performance measures is crucial - they need to be published with indications of uncertainty to communicate the sensitivity of the reported measure.

5. Explanations of the limits of the composite may help with its interpretation and also with making the process more transparent so that it can be clear what policy objectives are being maximised. This may also make the results more acceptable to participants and may make the reward and penalty schedule attached to the composite, more palatable.

Notwithstanding the importance of considering these important methodological and policy issues, some pragmatism in the approach to composites may be appropriate. Often the construction of composites that are less than ideal, may nevertheless lead to important empirical and policy analyses, the search for better analytical methods and improvements in data capture.

This report highlights the issues which need to be taken into account in the construction of robust composite indicators so that they can be designed in ways which will minimise the potential for producing misleading performance information. If such issues are not addressed, composite measures may fail to deliver the expected improvements in performance or may even induce unwanted side-effects.

6. References

Audit Commission (2003a) *Technical manual on the use of performance indicators in CPA*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/TechnicalManualUsePIsCPA.pdf

Audit Commission (2003b) *Technical manual on the use of performance indicators in CPA: update for 2003/04 and 2004/05*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/CPASTCCTechnicalManualonPIs.doc

Audit Commission (2003c) *Comprehensive Performance Assessment Guidance for district councils Version F1.1*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/april1.pdf

Audit Commission (2005) *Service assessment framework: Technical guide to CPA 2005 for single tier and county councils*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/Nov05ServiceAssessmentFrameworksTG.doc

Benefit Fraud Inspectorate (2003) *CPAs of single tier Authorities*, BFI: Harrogate. http://www.bfi.gov.uk/about/products/cpa/local_authority.asp

Commission for Health Improvement (2003a) *NHS performance ratings acute trusts, specialist trusts, ambulance trusts 2002/2003*, Commission for Health Improvement: London. http://www.chi.nhs.uk/ratings/

Commission for Health Improvement (2003b) *Rating the NHS: A question and answer guide to the Finsbury Rules, How CHI CGR scores will affect 2003 star ratings*, Commission for Health Improvement: London. http://www.chi.nhs.uk/eng/ratings/finsbury_rules.shtml

Commission for Social Care Inspection (2003) *The Commission for Social Care Inspection*, Commission for Social Care Inspection: London. http://www.doh.gov.uk/csci/

Cutler, T. (2002) Star or black hole? *Community Care*, 30 May 2002, 40-41.

D’Agostino, R.B., Balanger, A. & D’Agostino, R.B.Jr. (1990) A suggestion for using powerful and informative tests for normality, *The American Statistician*, 44(4): 316-321.

Department of Health (2000) *Quality and Performance in the NHS Performance Indicators: July 2000*, Department of Health: London. http://www.doh.gov.uk/nhsperformanceindicators/indicators2000.htm.

Department of Health (2001) *NHS Performance Ratings: Acute Trusts 2000/01*, Department of Health: London. http://www.doh.gov.uk/performanceratings/2001/index.html

Department of Health (2002a) *NHS Performance Ratings and Indicators: Acute Trusts, Specialist Trusts, Ambulance Trusts, Mental Health Trusts 2001/02*, Department of Health: London. http://www.doh.gov.uk/performanceratings/2002/index.html

Department of Health (2002b) *Performance Ratings Methodology - Acute NHS Hospital Trusts*, Department of Health: London. http://www.doh.gov.uk/performanceratings/2002/method_acute.html

Department of Health (2002c) *A guide to social services performance “star” ratings*, Social Services Inspectorate: London. http://www.doh.gov.uk/pssratings/wholedocument.pdf

Department of Health (2003) *How SSI assess performance*, Social Services Inspectorate: London. http://www.doh.gov.uk/ssi/performance.htm

DETR (2000) *Indices of deprivation*, Regeneration Research Summary Number 31 Department of the Environment, Transport and the Regions, HMSO: London. http://www.odpm.gov.uk/stellent/groups/odpm_urbanpolicy/documents/downloadable/odpm_urbpol_021680.pdf

DLTR (2001) *Strong Local Leadership – Quality Public Services*, CM5237, Department for Transport, Local Government and the Regions, HMSO: London. http://www.europartnership.com/Documents/wp_part1.pdf

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

Hauck, K., Rice, N. & Smith, P. (2003), The influence of health care organisations on health system performance, *Journal of Health Services Research and Policy*, 8(2): 68-74.

Healthcare Commission (2004), 2004 performance ratings, Healthcare Commission: London. http://ratings2004.healthcarecommission.org.uk/

Healthcare Commission (2005), 2005 performance ratings, Healthcare Commission: London. http://ratings2005.healthcarecommission.org.uk/

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.

Kmietowicz, Z. (2003) Star rating system fails to reduce variation, *British Medical Journal*, 327: 184.

Miller, N. (2002) Missing the target, *Community Care*, 38.

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.

ODPM (2003) *Local government performance, Best Value Performance Indicators*, HMSO: London. http://www.bvpi.gov.uk/pages/index.asp

Office for Standards in Education (2002a) *The development of the Comprehensive Performance Assessment framework for the education sector: Briefing paper for Chief Education Officers*, Ofsted: London. http://www.ofsted.gov.uk/publications/docs/3021.pdf

Office for Standards in Education (2002b) *York comprehensive performance assessments*, Ofsted: London. http://www.ofsted.gov.uk/reports/manreports/642.pdf

Office for Standards in Education (2003) *Inspecting schools: Framework for inspecting schools (Effective from September 2003)*, Ofsted: London. http://www.ofsted.gov.uk/publications/docs/3266.pdf

OFSTED and DfES (2002) *Guidelines for the Education Profile of CPA: The contents and operation of the education profile for the 2002 Comprehensive Performance*

Assessment, Office for Standards in Education: London. http://www.ofsted.gov.uk/publications/docs/3020.doc

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.

Snelling, I. (2003) Do star ratings really reflect hospital performance? *Journal of Health Organization and Management*, 17: 210-223.