Conference on 'Where Does Britain Rank? International Public Service Rankings', Tuesday 13 December 2005, at One Great George Street, Parliament Square, London, 10am – 5pm

Reported on BBC Radio 4's 'More or Less' programme, Thursday 12 January, 2005. Please see **http://news.bbc.co.uk/1/hi/programmes/more_or_less/4604992.stm#email_form** for details, or **http://news.bbc.co.uk/media/audio/41210000/rm/_41210120_moreorless_12_01_06.ram** for the audio file.

EXPERTS FIGURE OUT RANKINGWORLD

What can be said about international rankings when a nation's position can move dramatically according to the methodology employed? Or, if France won a world health ranking contest in the year 2000, placing it first on a league table of 191 countries, and in the same year won the World Cup, could there be a correlation between the health of a nation and the strength of its national football team? And is a 'data clubs' approach to comparative statistics the only way forward in producing worthwhile performance rankings?

These questions and many others were raised and discussed in this conference on international rankings. Christopher Hood called it 'rankingworld' – the current preoccupation in media, academia and government with performance rankings of various kinds, from health service effectiveness, to educational performance, to worldwide transport efficiency. Christopher Hood's opening paper introduced a 'grand' league table of league tables, detailing Britain's place relative to 12 other advanced countries; Britain's place was third from bottom, above France and the United States. Having introduced it, however, Christopher Hood went on to highlight the challenges involved in generating the table and why it should be approached with a good deal of scepticism, which served as the subtext for the speakers who followed.

The World Health Organisation's year 2000 health ranking was criticised as a particularly questionable example of a world ranking, while an OECD educational performance league table, which placed the UK 7th and Germany 20th, came under attack from Alison Wolf for producing a league table that compared diverse educational systems and could look wildly different with slightly altered data interpretation. Good news for 'rankingworld' emerged, however, with word of a 'data clubs' approach, which involved the voluntary participation of worldwide metro organisations and had produced successful results, and a robust defence of 'rankingworld' from Geoff Mulgan, who described it as a useful tool in holding governments in 'arrogant countries' to account.

Christopher Hood characterised the mood of the meeting in his closing remarks as one of 'practical scepticism'. The conference was funded jointly by the ESRC Public Services Programme and the CMPO, and brought together representatives from academia, central government and the media. Attendance was strong, with 15 speakers and 70 participants, representing a number of universities, government departments, focus groups and other organisations, from a total of 9 countries.

For a more detailed report of the conference, please go to **http://news.bbc.co.uk/media/audio/41210000/rm/_41210120_moreorless_12_01_06.ram**, or read below.

Programme:

Opening Session: *Background*

Background paper How Does Britain Rank and How do We Know? International Rankings of Public Service Performance' – Christopher Hood/Craig Beeston (Oxford)

A statistical perspective on international public service rankings – George Gaskell & Jouni Kuha (LSE)

Response – Christopher Pollitt (Rotterdam)

Second Session: *Crime, Education & Health*

Comparing Health Performance and the 2000 WHO health rankings experience –

Andrew Street (York)

Comparing Educational Performance and the OECD PISA rankings – John Cresswell (OECD)

Comparing Crime and Police Performance – Martin Killias (Lausanne)

Response led by Ted Marmor (Yale), Alison Wolf (King's) and Tony Travers (LSE)

Third Session: *Transport*

Transport Performance and the Data Clubs Approach – Steve Glaister and Richard Anderson (Imperial)

Response led by Tony Travers (LSE)

Fourth Session: *The Way Forward*

The Way Forward: Using and Developing International Comparisons of Public Service Performance

- Wendy Thomson (McGill, ex UK Cabinet Office)

- Nick Manning (World Bank and OECD)

Response – Geoff Mulgan (Young Foundation)

<u>Background paper</u> How Does Britain Rank and How do We Know? International Rankings of Public Service Performance' – Christopher Hood/Craig Beeston (Oxford)

Christopher Hood began by drawing attention to the proliferation of rankings data~~,~~ in what he called 'rankingworld', and the World Bank characterised as a 'virtual explosion of datasets'. He went on to make the following points:

1. Three main causes for this increase in performance indicators were:

   1.1. They attract media attention~~.~~ .

   1.2. They are studied avidly by public-service leaders, the same who claim such data are problematic~~.~~.

   1.3. They appeal to 'change agents' who wish to drive reforms of public services.

2. Britain's ranking was examined by means of a 'traffic-light analysis' of Britain's place relative to 12 other advanced states on 21 available indices of public service performance, placing Britain in the top, middle, or bottom third. Overall, Britain came third from bottom on this measurement, above France and the United States and narrowly behind Belgium, with Singapore, Japan and Sweden coming first, second and third respectively.

3. Many methodological problems lay behind the 'traffic-light analysis' that were endemic in any analysis of performance datasets. Among these were the following:

   3.1. The data could be read in more than one way –Britain's 'red' performance to some could mean that vigorous attempts to reform public services were failing, while to others it could mean that further radical measures were required.

   3.2. 'The messy middle': while there might be especially good or especially bad performers, the vast majority often sat very close together in the middle, with differences so small as to be more confusing than helpful.

4. Christopher Hood concluded:

   4.1. Tables of this kind were useful for a 'first pass' to 'explore overall patterns', but the problems involved, if accounted for, would demand a very complicated analysis indeed.

   4.2. Britain's international position depended on which ranking was used, and evidence was only robust where there were 'knockouts', or many data pointing in the same direction. Instances of this were few.

<u>A statistical perspective on international public service rankings</u> – Jouni Kuha (LSE)

1. Kuha used two examples, the UNDP's Human Development Index (HDI) and Transparency International's Corruption Perceptions Index (CPI) to illustrate the distinction between objective and subjective indices. The HDI was *objective* because it was based on official data, and the CPI was *subjective*, or based on opinions and perceptions.

2. These studies raised different methodological issues:

2.1. *validity*, or how well an indicator captures what it purports to measure.

2.2. *reliability*, or consistency of measurement~~;~~.

2.3. *replicability*, or how transparent and reusable is the data.

3. Two particularly important properties of useful performance indicators were transparency, which improved an indicator’s perceived legitimacy, and comparability across cases and across time. This fostered the kind of knowledge necessary to improve performance.

4. All measurement indices had errors and uncertainties that needed to be understood and appreciated. Thus, measurement data should not be taken at face value, without knowing their origins and methodology.

<u>Response</u> – Christopher Pollitt (Rotterdam)

The conference topic had the 'aroma of a coming thing', but Christopher Pollitt confessed his desire to ‘throw a few small spanners into the wheels of the bandwagon before it rolls too far’. His main question was: 'who is supposed to use this data, and to inform precisely what kind of decisions?' He outlined several key issues:

1. What is the data being used for?

1.1. A supply and demand paradox emerged in that there was a huge demand for performance data, despite its acknowledged faults: ‘to put it bluntly, there’s a huge demand for duff data’.

1.2. Not one of the three main purposes for high demand was particularly noble, and one was not a purpose at all.

1.3. Christopher Pollitt posed the query: ‘are we to believe that hard-nosed chief executives in the private sector were likely to use these data as a guide to major investments?’ His answer was: ‘my God I hope not’.

2. What is the position of the UK?

2.1. Britain’s strong ranking on general governance, as compared to its ranking on specific public services, could be explained by a gap between rhetoric and delivery. The way down from policy to implementation was a long one, suggesting that, as with any large organisation, central government could not control the periphery with any degree of precision.

3. What should measurement priorities be?

3.1. Precise, reliable data would help far more than aggregated rankings to improve services on the ground, and a longitudinal comparison of whether Britain was improving its performance relative to itself would be more useful than comparison with other countries.

<u>Questions</u>

Questions focused on the methodological problems that had been raised by the speakers. David Halpern suggested that the need was to look more carefully at the interrelations between indicators and the patterns they suggested, while Sir Christopher Foster expressed sympathy with Christopher Pollitt, suggesting two areas, crime and transport, where the UK's figures had been misleading. In response to criticisms of the validity of much of the data, Christopher Hood acknowledged that such problems were significant, and the aim of the background paper had been to review what data were available across the public services.

Second Session: *Health, Education and Crime*

Comparing Health Performance and the 2000 WHO health rankings experience –

Andrew Street (York)

Andrew Street was highly critical of the methodology behind the WHO's year 2000 world health rankings. He pointed out that 'econometric gamesmanship' had been used to produce a coherent analysis of the data.

1. Replicating the same methods used by the WHO, Andrew Street developed a parallel case study to demonstrate the dubious nature of the WHO's methodology:
    1.1. An analysis of the FIFA world football rankings, published at the same time as the WHO health rankings in 2000, were compared to the results of the WHO study using the same methodology.
    1.2. From this it was possible to draw the conclusion: 'if policymakers wish to improve the health of their nation, efforts should be made at strengthening the national football team'. This, of course, was 'absolute nonsense'.
2. The WHO's claim that its data was 'stable under numerous specifications' was untenable. The University of York undertook its own investigation into the WHO data, testing many apparently innocuous modelling decisions made by the WHO, and found that the results were highly sensitive to change and therefore not robust.
3. Nonetheless, two main positives had emerged from the WHO 2000 report:
    1.1. It had been an innovative exercise, introducing interesting new concepts in health measurement.
    1.2. It had stimulated a vast amount of debate beyond the normal circles, engaging academics and the public in a discussion of what health systems really ought to achieve.

Comparing Educational Performance and the OECD PISA rankings – John Cresswell (OECD)

John Cresswell introduced the OECD's Programme for International Student Assessment (PISA), with the following details:

1. PISA was an assessment of student ability in the 30 member countries of the OECD, undertaken from the year 2000.
2. The first test in 2000 had 28 participating countries, and focused on reading literacy. It was identical in all countries: a two hour test followed by a questionnaire administered to 175,000 15 year olds at school, from a sample of over 150 schools.
3. The UK scored above the OECD average, and came 7th in the ranking. Finland scored 1st, while a shock result put Germany outside the top 20.

4. Some interesting comparisons, analyses, and conclusions could be drawn from the data, proving they were useful not only in broad judgements but in a drilled-down, finer analysis:

   4.1. The UK’s percentage of students in the highest bracket was quite high, on which scale the UK ranked fifth.

   4.2. Many of the most successful countries in the ranking, including Finland, achieved their success by reducing the number of students in the lowest achievement bracket.

   4.3. A comparison of scores between reading, mathematics and science suggested that for some countries, such as Japan, where science and maths levels were very high but reading was not so high, the disparity was significant, while for others, such as the UK, no significant differences could be isolated.

   4.4. The data strongly pointed to gender differences. This was particularly true of reading, where females scored consistently better than males across the board, and true to a lesser extent with mathematics, where males were generally the higher scorers. Finland showed the greatest gender disparity, where female performance was extremely high, while the UK showed no significant difference.

   4.5. The questionnaire portion of the test was designed to assess the home background of students and what effect it might have on their performance. This was difficult to do in comparison, since measures of socio-economic status varied from country to country; however, the data produced indicated that in some countries home background had a higher impact than in others. Students in the UK, for example, were more likely than average to be affected by background, while students in Korea, Iceland and Finland were less likely.

Comparing Crime and Police Performance – Martin Killias (Lausanne)

Martin Killias argued that the uses to which data were put were as important as the production of the data itself. Rankings were not just instruments for theory testing or evaluation but also blame allocation~~;~~, and at that point~~,~~ they became less useful. Martin Killias drew on an example of a more ‘thoughtful’ crime indicator, the International Crime Victim Survey, which had been developed since 1993 as a reaction against the deficiencies of UN crime indicators, to show that data had value depending on how it was used:

1. The ICVS was a survey of crime victim experiences, beginning with 14 countries and then expanded. This allowed more meaningful experiential questions to be asked than Transparency International’s ranking, which put the stress on tiny differences between nearly similar systems. For example, in Western Europe, the difference between TI’s ranks on corruption was often marginal; rising or falling a few places was merely a result of ‘noise’ in the data. Despite its lack of real significance, the indicator caused Governments extreme concern about where they ranked.

2. Martin Killias was enthusiastic about the de Soto approach of reflecting government performance by measures of procedural experience, such as how long it took to replace lost documents, like a driver’s licence. In an extreme example, where an attempt to establish a bakery shop would take 300 line-ups and 2.5 years

of full time work, corruption was the only solution. This could be more revealing and valid than questions such as confidence in the police, which was culturally variable.

~~4.~~3.Overall, despite all the definitional and methodological problems, the important thing was to develop better data rather than respond to the inadequacies of the current data. The data was useful as a benchmark, or a way of challenging the 'most stupid policies' of government, if used correctly.

<u>Response</u> led Ted Marmor (Yale), Alison Wolf (King's) and Tony Travers (LSE)

Ted Marmor

Responding to Andrew Street's paper, Ted Marmor made the following points:

1. The York team's critical work on the WHO rankings was entirely accurate; however, Andrew Street's conclusion was inaccurate in suggesting that there was any value in the ranking whatsoever, as the methodology was so poor.
2. The problem lay not in the questions posed, but in the answers produced, and the purposes to which they were put. For example, the question of 'how fair' was the financing of health care never received an adequate codification: what was meant by fair?
3. Having previously challenged the intellectual originator of the WHO work with some of these problems, and received an assurance that the faults were known but that the study would be valuable in generating comments for the next time around, Marmor found the WHO work deeply inadequate and unnaceptable. The explanation did not count as justification; the WHO ranking could not be defensible on the basis of its generating controversy and media interest.

Alison Wolf

Alison Wolf furthered Ted Marmor's sceptical tone on the methodological problems behind performance measurement when she responded to John Cresswell's discussion of education rankings. She made the following points:

1. Education ranking on a global level entailed more methodological problems for several reasons:

    1.1. It involved imposing a single scoring system on countries that had extremely different education systems and goals for those systems. This was an irresolvable methodological problem with international rankings.

    1.2. Many measurement challenges were far from trivial, such as the issue of translation, which by nature precluded uniform answers.

    1.3. How a survey was to be sampled raised often violent debates, as it went to the heart of what the survey was trying to achieve. Questions of whether the sample should be by age or grade, for example, could make a massive difference, because some countries allowed repetition of a year while others didn't, depending on their objective for performance among their students. This could affect the sample of the study significantly.

2. Rankings had a huge impact politically. Politicians were often the users, because they were simple and easy to digest in a market flooded with complex data, and they could be used to prove the success of particular policies. However, little could realistically be deduced from comparative datasets.

Tony Travers

Tony Travers addressed the issue of why benchmarking exercises were being so widely undertaken, with the following explanations:

1. An underlying cause might be that nations were seen as competing, and rankings in health, education, and crime were indicators of international economic performance. Fuelled by the fear that a country might fall behind economically, rankings were seized upon as a way to drive up performance and therefore economic growth.
2. Three initiatives lay behind the work:
   - 2.1. International organisations seeking to justify their own funding.
   - 2.2. Organisations seeking to raise their profile – rankings got immediate media attention.
   - 2.3. Universities seeking knowledge, unpicking the poor quality of the work being done.
3. For four reasons, rankings mattered:
   - 3.1. They affected the quality of output in public services. Once produced, the data would be used to push reforms.
   - 3.2. They were used for getting more money into public services.
   - 3.3. They could be used to apply pressure on public services to be more efficient.
   - 3.4. They could be used by institutions to compare their performance with other institutions, in order to prove they are performing better.
4. The existence of cheap computing made the work much easier to perform; this had led to a mushrooming in the number of experts, consultants, and international organisations interested in the work.
5. Performance indicators should be supported in general principle, but there was a worry that benchmarking was very susceptible to distortion: having generated a great deal of political interest, data were often put under a significant amount of pressure.

Questions

Comments after the second session expanded on the reservations made by the speakers, but several speakers sought to moderate what was perceived as an excessively pessimistic tone. Christopher Foster drew attention to the ‘most criticisable factor’ – the fact that overly simple performance indicators were put to use as incentives and target setters. Alison Wolf expanded on this point by querying why governments felt the need to control every particular of, for example, the education service, and how they found it possible to believe that an effective set of incentives for the smallest levels of education could be developed.

Frank Vibert expressed reservations about the policy implications of some rankings, particularly composites whose components did not necessarily cohere in ways that everyone would accept, and he cited the Human Development Index as a case in point. He worried about the selectivity that led to some measures being used and others ignored, thus producing an incomplete picture. He argued, however, that even less than perfect performance data were preferable as a basis for allocating funds (for instance by international bodies) than the alternative: governments and international groups trying to determine policy 'in smoke filled rooms', by 'seat of the pants driving'. In response, Wolf argued that it was worse to make a decision based on bad information than none at all.

Third Session: *Transport & Health*

Transport Performance and the Data Clubs Approach – Steve Glaister and Richard Anderson (Imperial)

Steve Glaister and Richard Anderson introduced a consultancy based at the Engineering Department of Imperial College called the Railway Technology Strategy Centre, which approached benchmarking in a unique way. Those supplying data for analysis, and those interested in the results, were voluntary members of a 'club', and as such they were able to share their data frankly and freely. They highlighted the following issues:

1. The first 'group approach' of this kind began in 1982, when the London Underground and the Hamburg Metro began comparing the productivity of their organisations. This led to the creation of CoMET, the Community of Metros, in 1995, a group of 5 of the world's largest metro companies which expanded to 12 by 2005.
2. The approach had been highly successful, for the following key reasons:

    2.1 Clarity of purpose. Data collection and analysis was not a means to itself, but a means to delivering benefits to the participants. There was a key performance indicator system (KPI) at the heart of the forum, but it was used to stimulate questions, such as *why* one metro system might be doing better than another.

    2.2 Longevity. The work had been ongoing on an annual cycle for 11 years, which had a significant impact on effective analysis, as many answers took years to emerge.

    2.3 Confidentiality. An agreement existed between members of the club that fostered the free flow of internal data between the organisations.

    ~~2.4~~Direct contact. Senior members of the CoMET companies met face-to-face twice yearly to discuss points of comparison. This was an extremely important element, though it demanded a limit on the size of the group to 12.

    ~~2.5~~2.4 Ownership. Each participant in the club drove the work, meaning it was not overly theoretical, and was ultimately about implementation and best practice.

3. The trust engendered by the club approach allowed for some interesting results: for example, Asian metros were up to a hundred times more reliable than European metros. A 9-5 commuter could travel on the Hong Kong metro every day for one year and only be delayed by 5 minutes once, in significant contrast to the experience of most European and North American metros.
4. The sensitivity of key performance indicators allowed for what was within management control and what was not, such as local labour laws, helping managers determine what was within their power to change.
5. Success in the process of benchmarking existed in persuading an organisation to ask the right questions of itself, which the data clubs approached had managed to do during its history.

Response – Tony Travers

Tony Travers highlighted the implications of the data club approach for all other kinds of benchmarking activities:

1. The enormous importance of shared expectations and drive to deliver among those involved.
2. The voluntary nature of the club gave the members ownership of the process, so that participants were not being asked for data that might later be used against them.
3. The independent technical handling by Imperial College was an important element in that it removed the data from supposed invested interests.
4. International ranking did seem the appropriate level of comparison in the case of metros, where no natural comparator of equal circumstances – such as size and cost – tended to exist within the country.
5. Among the issues raised by the data clubs approach, there emerged a paradox: these were public sector institutions using public money that could only work effectively with data because they were kept secret. This implied an interesting trade-off between transparency and effective participation.

Questions

The data clubs approach met with significant interest, and many questioners sought further details as to how the operation was run, established, and why more had not been heard of it. Richard Anderson indicated that in many ways the process had snowballed, having begun as an inward-looking enterprise when the Managing Director of London Underground began a practical inquiry into London Underground's efficiency. As the approach proved robust over time, new uses were found and interest grew.

Other questions focused on how 'generalisable' the approach was, or how far it was applicable to other industries, such as the postal service. Tony Travers felt that there were many commonalities that could be read across to services other than transport, while Steven Glaister reemphasised the need in all cases to include biannual meetings, where members could meet and dictate to what uses the data should be put.

The Way Forward: Using and Developing International Comparisons of Public Service Performance

Wendy Thomson

Wendy Thomson stressed the usefulness of international comparative data, if not rankings themselves:

1. There was still a big gap between what people thought constituted a successful public service and what some of the more data rich analyses said. Complexity was unnecessary: the most useful data were often not very sophisticated at all; school league tables, for example, were much criticised for over-simplicity but were compelling at same time.
2. While composite rankings might be less useful in some circumstances, less aggregated indicators could often be valuable. Rankings did serve a role for people who wanted to make services work.
3. It was difficult to determine what kind of international comparative data would be most useful. It would take a huge amount of effort to make the data sound, but several approaches would signify improvements. For example, England's Audit Commission's Comprehensive Performance Assessment approach, if it could be made practicable in a world setting, would be valuable, while the rankings themselves should combine administrative data as well as informed professional judgements, keeping some level of perspective.
4. To make data useful, it was important to ensure the public could understand and make sense of it. Individuals needed to be able to validate it from their own experience; nobody could be told, for example, that something was true from a national perspective if that figure was not verifiable from individual experience. Public interest lay not in big numbers but in specifics, so it would be important to make data useful by applying it small-scale. A parent might not wish to know the state of education as a whole, but rather how a particular school, or even teacher within that school, was performing.
5. Academics could act as a 'parallel source of intelligence' on international rankings, serving to expose those of poorer quality and to act as a stabilising influence on the 'media frenzy' that attended new publications.

Nick Manning

Nick Manning discussed what he called 'our stumbling approach': the OECD's attempt to produce exploratory papers on governance in OECD countries. He raised the following points:

1. They operated in a world of weak comparative data. There existed a lack of data on public administration that meant there was little of real use to say in comparing one country to another. In response to this data gap, the OECD was producing 'at a glance' papers, essentially 'stapled together' aggregations of everything known about the topic.
2. The OECD's own data, gathered for the first of the publications called *Governance at a Glance*, was 'pretty shaky' in some areas. A very significant

amount of the data failed any reasonable test of reliability, turning to ‘mush’; the magnitude of the cull that took place in finding what data were useful was vast. The papers would have something to say, therefore, but would be relatively thin publications.

3. There were two impulses behind the push of international organisations towards doing this kind of work:

    4.1. Entrepreneurs. This group of analysts saw a problem in the lack of globalised ‘advice’, and sought to make good ideas transferable. Their impulse was to cut through complexities and produce simple statistics that would rank countries easily, thus encouraging failing countries to step up.

    4.2. ‘Trainspotters’. These individuals were simply obsessed with their own topic and wanted to push out as much data on it as possible. They enjoyed ‘wallowing around’ in the data, rather than producing simplistic figures to drive performance or policy.

5. A problem existed for the OECD effort as much as any other in that it was necessary to get funding for the work. Member governments were pushing very hard for rankings, possibly in the hope that they would be ranked first in some dimension. It was hard to resist this pull; the idea of ‘slowly building up datasets’, which would be the best way of ensuring the soundness of the data, came across as ‘unsexy’ and painted data collectors as trainspotters, which did not encourage the allocation of funding.

6. It was almost accepted within the OECD that there would be a push towards rankings sooner than was best for the data. The question remained open whether the OECD could resist the pressure; it was possible, though, that the data clubs approach might offer an appropriate outlet.

Response – Geoff Mulgan (Young Foundation)

Proclaiming himself to be a ‘moderate supporter’ of the proliferation in comparative data, Geoff Mulgan quoted Margaret Thatcher as having suggested to a civil servant that what was needed was not better answers, but better questions. He disagreed with Alison Wolf and Ted Marmor in his belief that some questions, even if distorting, were better than none at all.

1. The proliferation in datasets could not be traced to media pressure. It was more the result of non-governmental organisations and public trends; moreover, governments found them useful, so journalists were, in this sense, intermediaries and not drivers.

2. Datasets were useful for holding arrogant countries to account, and most countries were arrogant. The English speaking world was especially arrogant, with governments better at publicising results than producing them; performance data served as an ‘antidote’ to this impulse.

3. Usefulness also existed in locating connectedness between data, from which persuasive hypotheses could hopefully be made, and action would result.

4. Several issues existed that affected the strength of datasets, to which there were no answers yet available:

4.1. Regularity. Comparison was often misleading if it involved wildly different sized countries. It was always necessary to ask: is this the right scale of regularity? A national-level comparison was hardly ever the correct scale.

4.2. Path dependence. Snapshots were useless; a historical sense was needed for useful analysis, in order to show how performance had changed over time.

5. For the future, more accumulated datasets were not needed, but rather the intelligent seeking out of patterns within those datasets. This was not the skill of statisticians, but would come as a result of the cross pollination of statistical experts and those asking the questions, as had happened with the conference today.
6. For all their flaws, rankings and comparative data exercises combated complacency, gave rise to new questions, and provided ladders for ascending to new levels of understanding that could be discarded once gone up. Critics of datasets should remain silent unless they had a better solution.

Questions

The issue of whether a more ‘descriptive’ approach to performance assessment would be more effective was raised, following from Martin Killias’ earlier suggestion that corruption was an issue not properly understood from raw data. Nick Manning agreed that this would be an effective approach, so long as the questions asked could be unified and analysed effectively, otherwise Geoff Mulgan’s suggestion that rankings were ‘dearrogantising’ would no longer be true. Martin Killias agreed it was important to ask the right questions – defined, clear and practical – while Geoff Mulgan agreed that data were often tools of the state, not the citizen, due to an assumption that statistics implied a ‘truer truth’ than personal experience. If it was true that people were better at measuring their life experiences than were statistics – for example, how long they had to wait for an operation, as a basis of measuring the effectiveness of the health service – more effort should be made to include personal experience and exclude abstractness. Wendy Thompson agreed that a ‘compelling story’ would be a stronger impetus for change than abstract data.