6 Measuring Success

> We are all benchmarkers now.
>
> Jacques Santer, former President of the European Commission (quoted in Sisson et al., 2003, p. 16)

In the struggle to create better forms of public services it is important to remember that there will never be "perfect" models of service delivery. Nor will public services look exactly the same across time, place or sector, with public preferences and priorities shaped by their social, economic and geographic contexts.

And yet, it is essential that we are able to measure the performance of public services across these different dimensions to assess the degree to which they are meeting their stated goals of reform – particularly the extent to which they have moved beyond the logic of marketization. Without some widely agreed upon form of measurement it is difficult to learn from and share "good practices" in the building of a global pro-public movement.

Unlike some critics of performance evaluation, I will argue in this chapter that the benchmarking of public services is an important and productive component of pro-public reform. It is also unavoidable, given that standardized benchmarking systems already exist in virtually every service sector in every corner of the world, with well-established and powerful adjudicating agencies determining what constitutes success and failure. In other words, performance evaluation is already being done, much of which measures, celebrates and reproduces commercialization tendencies, necessitating some form of engagement with this enormously influential trend in public services. Relatively little critical attention has been paid to the origins of these benchmarking systems, their lack of suitability for assessing progressive pro-public reforms, and what can be done to change them.

My goals in this chapter are threefold. First, to demonstrate the growing influence of benchmarking in public services and its bias towards top-down, market-oriented and Eurocentric indicators. Second, to impress upon readers the value of benchmarking despite these problems. And third, to outline a proposed new model of public services performance evaluation that could

DOI: 10.4324/9781003293002-8

be adopted in different sectors in different locations in ways that are more democratic and less commercial than the benchmarking systems currently in place. This proposed model follows the same principles outlined in previous chapters as they relate to developing new conceptual frameworks of publicness – i.e. a core set of universal normative values that are broadly progressive and less-commodified than existing "public" norms, while at the same time allowing for flexibility in terms of how new standards are assessed and weighted in different contexts. My focus here will be on the water sector (which I am most familiar with) but the arguments are relevant to other public services, all of which could (and should) develop progressive new forms of benchmarking with similar principles of universality and flexibility.

As with all of the changes proposed in this book, altering existing forms of performance evaluation systems in public services will not be easy. There will be stiff resistance from agencies that have invested heavily in current benchmarking systems (including the World Bank and the International Organization for Standardization). There will also be resistance from public service managers who have devoted considerable time and energy into developing these models and adopting them into their daily routines (particularly managers who benefit from marketized indicators and whose salaries are increasingly tied to achieving or maintaining certain performance standards). Even managers and policy makers who do not like existing benchmarking systems, and who may be committed to the idea of new, pro-public performance measurement systems, will find it difficult to shift gears given the enormous time and energy required to develop and implement new models, particularly if they include the sort of contextualized flexibility I am proposing here. The fact that most performance evaluations are conducted by technical staff (such as accountants and engineers with little qualitative or political training) exacerbates the challenge.

The Rise (and Rise) of Benchmarking

Modern benchmarking originated with private industry in the 1950s and was popularized by the Xerox Corporation in the 1970s in the United States in response to what the company saw as a rising competitive threat from Japanese technological firms (Camp, 1989; Levy & Ronco, 2012; Sisson et al., 2003). Benchmarking is grounded in the use of quantifiable performance indicators for analyzing the internal activities of individual organizations, but takes on comparative yardstick characteristics when used to compare performance across organizations as well as across place and time (Pidd, 2012). For benchmarking to work, therefore, organizations must all (and always) use the same performance metrics. By the mid-1990s "almost four out of five companies in Europe, North America and South East Asia were reported to be using benchmarking", prompting then-president of the European Commission, Jacques Santer, to claim that "we are all benchmarkers now" (Sisson et al., 2003, p. 16).

Enthusiasm for benchmarking in the public sector developed soon after, concurrent with (not coincidentally) the rise of New Public Management and the push for privatization. In fact, it was the privatization of public services in the United Kingdom during the 1980s that was arguably the single most important catalyst in the adoption of benchmarking by public agencies, with the creation of the Water Services Regulation Authority for England and Wales (OFWAT) in 1989, for example, intended to quantitatively track the effects of privatization over time. Regulators and policy makers in other countries took note, with the benchmarking process quickly spreading around the world. In the United States, the 1993 Government Performance and Results Act "precipitated a virtual orgy of measurement" in the public sector (Pollitt, 2000, p. 120).

So too did academics jump on the bandwagon, with an associated proliferation of scholarly work on the topic. Dozens of books, hundreds of articles, and entire academic and managerial journals are now dedicated to performance evaluation and benchmarking in a wide range of sectors (although remarkably little of it engages critically with the theoretical presumptions or institutional foundations of mainstream benchmarking systems (Francis & Holloway, 2007)).

Benchmarking has now been adopted in virtually every major public service. It is "widespread" in the energy sector, for example, with electricity distribution having "witnessed a wave of regulatory reforms aimed at improving efficiency through benchmarking … namely measuring a company's efficiency and rewarding them accordingly" (Farsi et al., 2007, p. 1). Similar dynamics have unfolded in health care, waste management, and a host of other services, most of which employ similar statistical evaluative methodologies recommended by centralized agencies (Resnick et al., 1995, Jamasb and Pollitt 2000; Ettorchi-Tardy et al., 2012; Leal Filho et al., 2016; Sharma & Balachandra, 2015).

The water sector has one of the most advanced and widespread benchmarking networks. After OFWAT's adoption of performance tracking in the UK there was a flurry of activity to establish international benchmarking systems, including the creation of the International Benchmarking Network for Water and Sanitation Utilities (IBNET) by the World Bank in 1996, the formation of task groups within the International Water Association (IWA) to explore "best practices" in the late 1990s, the publication of the International Organization for Standardization's (ISO) 24510 series on "Activities relating to drinking water and wastewater services" in 2007, and the establishment of an annual IWA-sponsored international conference on performance measurements in 2008 (Alegre et al., 2008; Bowerman et al., 2002; Cabrera et al., 2010; Cabrera et al., 2011; Danilenko et al., 2014; Parena et al., 2002).

There are now dozens of national water benchmarking associations and a growing number of regional groups. European water operators have been particularly active in this regard, with several established regional

benchmarking programmes in place (e.g. the European Benchmarking Cooperation, EBC (largely Dutch and Scandinavian) and Aquabench (largely German)). Additional programmes are being developed by the Organization of Economic Co-operation and Development (OECD), the European umbrella association for national water federations (Eureau), and the European Commission (which, unlike other models mentioned here could become mandatory since the Commission has full legislative powers).

Notably, there are very few benchmarking associations dedicated to water services in Asia, Africa or Latin America (Berg, 2013; Berg & Corton, 2007; Corton & Berg, 2009; GWOPA, 2009). Performance evaluation is widely practiced in these regions, but benchmarking systems are generally imported (some would say imposed) from organizations based in the North, often at the request of international financial institutions, and sometimes as part of loan conditionalities. The World Bank, for example, makes its IBNET evaluation system available to national and local governments that want to "receive financing for capital improvements" (Van den Berg & Danilenko, 2011, p. 4; see also WSP, 2010).

There also appears to be an (unstated) race to expand these Northern-based benchmarking systems to countries in the South. IBNET has declared itself to be "the first global benchmarking standard … providing a global yardstick against which utilities and policy makers can measure their performance" (Van den Berg & Danilenko, 2011, p. 2). Not to be outdone, AquaRating (a proprietary, for-profit benchmarking agency developed by the Inter-American Development Bank) claims to be "the only international system that facilitates an objective and comprehensive assessment of a utility's performance, making it a pioneering tool in the global water sector."[1] EBC meanwhile is piloting its benchmarking model in East Africa (on a not-for-profit basis).[2]

In short, performance measurement in the water sector is growing, and benchmarking frameworks are multiplying: "Performance measurement is here to stay" (Poister, 2003, p. 21) and "we will be hearing many new benchmarking stories in the future" (Cabrera, 2008, p. 7).

Benchmarking Consensus

Despite rivalry in the benchmarking community there is broad consensus amongst its advocates as to how and why it should be done. Most importantly, it is seen to enhance (and enforce) transparency and accountability amongst water operators by making performance data available to the public and allowing people to compare their water operator with utilities in other jurisdictions. It is also seen to create opportunities for public participation in decision making by allowing "customer groups and NGOs to exercise "voice" in an informed way"[3], as well as contributing to "consensus-based global solutions" for achieving the UN's Millennium Development Goals (ISO, 2012, 2). Finally, benchmarking is seen to provide a common language

and system for identifying and understanding "best practice", offering water operators the quantitative targets and concrete tools for getting there.

There is also broad consensus as to what gets measured. Although every benchmarking system in the water sector has its own unique features, most draw heavily (if not entirely) on the performance metrics established by the ISO 24500 series, with more than 260 standards for water quality, and another 550 related to water services more broadly (ISO, 2012, 3–4). The IWA, meanwhile, acts as a *de facto* gatekeeper of performance criteria by dint of its role as a global umbrella organization for water operators (public and private), especially since the creation of its "Specialist Group on Benchmarking and Performance Assessment" in 2010.[4] As a case in point, the EBC claims to be "fully aligned" with IWA and ISO benchmarking protocol, using these indicators as "repositories" of performance criteria "for reasons of standardisation" (EBC, 2014, 5). In other words, there may be a growing number of benchmarking organizations, and they appear to be in some form of competition with one another, but they operate on essentially the same principles and use mostly the same performance criteria.

It is not possible to list the 260+ plus indicators that commonly make up water service evaluation systems here, but the following are indicative of the type and range of measurements that take place: number of water and sanitation workers per 1,000 connections; length of transmission and distribution mains renovated; percentages of unaccounted-for water; number of complaints due to water supply interruptions; book value of fixed assets; number of microbiological tests carried out for various chemicals; volume of electricity consumed; total capitalized cost of self-constructed assets; per capita consumption of water; number of mains failures; average time to complete repairs; and average water charges for non-residential consumption.

Measuring Is Difficult

Consensus on how and why to do benchmarking aside, the actual collection of data is anything but straightforward. Even the best-trained and well-resourced of water operators find it difficult to stay on top of what has become an onerous process of data collection. The situation is much worse for under-capacitated utilities in low-income countries where it can be a challenge to gather even the most basic of statistics (Alegre et al., 2000, 2006, 2008; Berg & Corton, 2007).

Once collected, analyses and comparisons of data are no simple matter either. There are intense debates within the benchmarking community – albeit largely impenetrable to non-specialists – about the use of metric versus process benchmarking (Cabrera, 2008), partial indicators versus total or combined factor analysis, and differentiated statistical methods for analyzing complex information with linear programming methods such as data envelope analysis and stochastic frontier analysis (Abbott & Cohen, 2009;

Parsons, 2002). Different clusters of indicators are also used, with EBC, for example, employing five key "performance areas" (EBC, 2014, 7), IBNET using 13 categories[5] and AquaRating (2014a) using an entirely different set of eight "rating areas", ironically making a comparison of different benchmarking frameworks difficult.

As a result, abstract and highly technical forms of measurement take on a life of their own, often disconnected from larger utility objectives, yet shaping policy and practice. Externally-defined performance indicators can thereby skew in-house priorities, stifling local identity and resulting in an abstract fixation on outside procedural norms: "At its worst, instead of being a force for change, benchmarking can put a stop to serious analysis of problems and/or experimentation with [locally-defined] innovative solutions" (Sisson et al., 2003, p. 23). It can also result in an anxious feeling of "keeping up with the Joneses" (Valcik et al., 2012), even when high-level benchmarking may not be feasible or appropriate for a local water operator. In some cases, benchmarking can result in a situation where water operators are measuring for measuring's sake, with benchmarking becoming an end, rather than a means, to improved water services (Pidd, 2012).

The truism that "What gets measured gets done" (Osborne & Gaebler, 1992, p. 146) can aggravate the situation further, focusing attention on criteria that may not be the most relevant to a particular water operator or its end users. As Francis and Holloway (2007, p. 177) note, "[b]enchmarking is no more immune to the GIGO (garbage in, garbage out) principle than any other performance management system".

Criticisms of Benchmarking

Most proponents of benchmarking recognize these challenges but do not consider them to be fatal to the measurement enterprise. For the most part they are seen simply as another technical challenge to be overcome in the pursuit of standardized performance evaluations.

More radical criticisms of benchmarking do exist, but the literature is surprisingly thin and under-theorized given how pervasive the practice has become (Francis & Holloway, 2007, p. 172). I outline three critiques below, drawing as much on the work of those who celebrate benchmarking as I do on its detractors, weaving together an admittedly eclectic set of theoretical concerns related to commercialization, a lack of democratic process, and the belief that all forms of benchmarking necessarily impose a singular worldview on a diverse set of international beliefs and practices.

Promoting Commercialization

For proof that benchmarking promotes commercialization one need look no further than the literature from benchmarking advocates themselves. For many proponents of benchmarking in the water sector it is an intentional

and explicit tool for commercialization, introducing competitive pressures to an inherently monopolized sector by offering an "alternative to market forces" (van Helden & Tillema, 2005, p. 339). With little or no possibility of direct competition, benchmarking is seen to simulate and stimulate market behaviour, pushing water operators to lower costs and improve services. According to IBNET:

> Inter-utility performance comparison is needed in the water and sanitation sector because the sector offers limited scope for direct competition. Firms operating in competitive markets are under constant pressure to outperform each other. Water utilities are often sheltered from this pressure, and it frequently shows: some utilities are on a sustained improvement track, but many others keep falling further behind best practice. … Only the most efficient, financially viable utilities are able to respond to urban growth, connect the poor, and improve wastewater disposal practices.[6]

"Financial viability" is key here – a metric that many see as "the ultimate value of utility benchmarking" (Van den Berg & Danilenko, 2011, p. 8). Improving cost recovery and reducing per unit expenses have become the gold standard of measurement in the water sector, with financial indicators such as "percentage of unpaid-for water" or "number of employees per 1,000 connections" often being used as proxy for overall performance (see e.g. World Bank, 2014, p. 129). Financial criteria are not the only measurement standards but they constitute a large proportion of benchmarking indicators and attract a disproportionate share of attention from policy-makers and funders, reflected in part by the massive literature on financial outcomes in the water sector and support from development institutions such as the World Bank for improving cost recovery (Alexander, 2005; Breen & Doyle, 2013).

Some benchmarking advocates even see it as a way to promote privatization, forcing public water utilities to make their fiscal data available for corporate review and helping to identify markets for potential private investment. According to IBNET, "[p]rivate investors interested in expanding their interests in the water and wastewater sector can use [benchmarking] to carry out an initial screening of potential target utilities", helping them to "pinpoint those with revenue-generating potential" [7] and to "identify viable markets and opportunities for creating value" (Van den Berg & Danilenko, 2011, p. 4).

As such, involvement in the development and implementation of benchmarking criteria is often touted as an opportunity to be at the cutting edge of competitive water markets. The ISO encourages firms to "get involved" in the development of benchmarking because it "can bring significant advantages to your business" via the establishment of standards that fit with an organization's strengths or by giving advance notice of what the market

will expect. By creating market-friendly indicators, benchmarking can help firms "access new markets" and "facilitate free and fair global trade".[8]

Not surprisingly, many international benchmarking organizations are composed heavily of multinational corporations keen to shape global standards across a wide range of sectors, from environmental sustainability to corporate governance (Acuto et al., 2021; Bruno, 2009; Clapp, 1998; Nadvi & Waltring, 2004; Prakash & Potoski, 2006; Stevenson & Barnes, 2001). The ISO has come under particular fire in this regard, with critics arguing that most of its committee work is conducted in a handful of countries in the North and dominated by large private firms, making it little more than a "corporate private regime" (Haufler, 2004, p. 126). Proposals by the European Commission to develop an EU-wide benchmarking system for water have received similar criticism, with Aqua Publica Europea noting "the unbalanced nature of [the planning group's] membership" and the fact that "the public sector is scarcely represented, if at all" (APE, 2014, 6).

The result is performance evaluation systems in the water sector that "strongly motivate [operators] to be efficient and innovative, mitigating their operating costs and expenses" (Marques & Simões, 2010, p. 15; see also Bowerman & Ball, 2000). In this respect, it can be argued that benchmarking has contributed to the naturalization of financial efficiency in the water sector, converting economic goals into "neutral facts" that validate and reproduce otherwise contestable and diverse aims (Boelens & Vos, 2012, p. 18). Benchmarking can, as with any other "absolutized efficiency calculus", be used "as a weapon to suppress contending social groups, their social analyses, and their programs for social change" (Wolff, 2002, p. 3).

None of this is to say that benchmarking is inherently commercial. Nor does it automatically bind water operators to their financial bottom line. Many public sector water managers and policy makers are intensely aware of – and often opposed to – its commercializing influences. There is also little evidence that benchmarking actually results in improved financial discipline (Braadbaart, 2007). But given the influence of organizations such as the World Bank in promoting financialization and marketization in water (and other) services (Bayliss, 2014; Hunter & Murray, 2019; Loftus & March, 2015, 2016; Loftus et al., 2019), and their stated intent of using benchmarking as a market substitute, there are ample grounds for concern around the impacts of this growing trend on water operator behaviours and outcomes. The fact that none of the major water benchmarking systems mentioned above make any distinction between public and private water operators only serves to heighten these apprehensions.

Undemocratic Processes

A second concern with benchmarking is that it tends to happen in a top-down, undemocratic manner, excluding service users and lower-level employees. Because of the highly technical nature of performance evaluation most

benchmarking systems make little (if any) effort to involve non-experts. Most water users are completely unaware of benchmarking processes and the 260+ performance indicators that can go into them (let alone the advanced statistical analyses that accompany their assessment), essentially rendering the benchmarking process opaque. Nor is the average water user likely be tempted to participate in such a highly specialized and mechanical procedure.

Large benchmarking organizations have done little to make these processes more transparent or inclusive of the broader public. Benchmarking reports are technocratic in method, difficult to decipher, and seldom available to the general public, although "simply publishing the results does not necessarily equate to more transparency" either (APE, 2014, 2). Access to IWA and ISO benchmarking models requires an expensive membership, while AquaRating is a proprietary framework (Krause et al., 2012). Little wonder that benchmarking is the domain of a relative few.

Some critics see in these actions a form of neoliberal governmentality, with monopoly control over a "governing technology that seeks to facilitate the self-governing capacities of individuals and/or organizations through the production of a normalizing knowledge" (Triantafillou, 2007, p. 836). Insofar as benchmarking is an attempt to create an "international consensus on solutions to water issues" (ISO 2012, 4), it is seen by some as a mechanism for "producing truth … steering social behavior and giving normative meaning to particular water practices of particular water user groups … convincing not only the actors who have to apply these concepts but also the creators themselves" (Boelens & Vos, 2012, p. 18; see also Vos & Boelens, 2014).

Benchmarking systems, critics argue, must be seen within the "social and power relationships" in which they are embedded, with planners and managers "often not aware of their value-loadedness, convinced that they provide objective advice that should be adopted by policy-makers and politicians" (Boelens & Vos, 2012, p. 24). In this regard, benchmarking can validate concepts of success and reinforce "best practices" that advance particular agendas. The truism cited earlier – "What gets measured gets done" – can result in an unremitting cycle of investment and policy emphasis that advances the commodification of water and justifies inequities, particularly in countries in the South where opportunities for democratic input and alternative voices are often very limited.

Once again, this neoliberal governmentality need not be an inherent feature of benchmarking. Nor is it always an intentional outcome on the part of public water managers or benchmarking designers – some of whom express concern with benchmarking's exclusive nature but have no other options to employ (APE, 2014). Nevertheless, the reality of current performance evaluations is that they are largely impenetrable to the general public, while the lack of robust debate on the topic is itself indicative of the enigmatic nature of benchmarking in practice.

Imposing Universality

A third criticism of benchmarking is that universal performance criteria essentialize and homogenize public services and the people that use them, running roughshod over cultural and political difference by imposing generalized standards on the world. The argument here is partly an epistemological one, questioning the validity of objectivity and universal standards in comparative performance evaluations. Critics argue that there are no constant, universal truths: "Democratic conceptions of the common good will always be partial and provisional, never universal or static … the common good can never be specified *a priori* as an input for the political system or as a static measure for the quality of governance" (Dahl & Soss, 2012, p. 31). For Zwarteveen and Boelens (2014, pp. 151–2), "knowledge about water will always and necessarily be uncertain and provisional," reminding us that we must "remain vigilant about the temptation to unequivocally use "science" and the objectification it entails in dealing with water's complexity." For these critics, it is simply not possible to measure and compare performance in public services across place and time because no two communities are ever alike, while public conceptions of water, education, health care and other services are constantly changing.

The implication here is that benchmarking can smother the very improvements it seeks to promote. In attempting to impose standardized measures of value, measurement serves to "fix the public in place, rendering it static in a manner that contradicts the promise of an evolving constitutive democratic process" (Dahl & Soss, 2012, p. 22). What is required instead, critics argue, is a focus on the creation of space and resources for "deliberative processes that … allow new conceptions of public identities and interests to emerge" (Dahl & Soss, 2012, pp. 22–3). Rather than relying on a notion of a public that is "always a pre-existing collectivity … that can be identified, addressed and moved to action," it is argued that we should see values as "entities that are always mediated and always emergent" (Mahoney & Clarke, 2013, p. 932). Relaxing the demand for universal standards "usefully opens the door to accepting diverse and plural knowledges about processes of water-related change – including those based on the experiences and knowledge of people who live in changing environments" (Zwarteveen & Boelens, 2014, pp. 151–2). Benchmarking is seen to disqualify and marginalize alternative forms of public service management and valuation (particularly "pre-modern" forms), taking on the hue of colonial practice "imposed from an outside position as a way to bound or police governance" (Dahl & Soss, 2012, p. 31).

These critics are particularly concerned that benchmarking standards have been developed by institutions in Europe and North America, with insufficient attention being paid to the realities of public needs and expectations in the global South. Cultural, political, economic and environmental differences may make some performance criteria irrelevant or inappropriate,

while the sheer cost of benchmarking can make it impossible or irresponsible to do in full. Even within the European Union there are concerns about the suitability and feasibility of universal benchmarking models, with Aqua Publica Europea arguing that "many operators in rural or less favoured areas face cultural and economic difficulties in participating in benchmarking exercises" (APE 2014, 6). This concern is magnified many times over in lower-income regions in Africa, Asia and Latin America.

Down with Benchmarking?

Given these critiques, should benchmarking of public services be abandoned? Yes and no. To the extent that benchmarking intensifies market pressures, excludes the general public and imposes inappropriate goals on diverse practices it can be argued that it is not suitable for many places. If nothing else, it is crucial that there be more research and a more robust debate on how benchmarking plays itself out in practice, particularly in countries in the South.

But I am equally convinced that benchmarking of some kind is essential to the advancement of "good" (as opposed to "best") public services. It may be problematic to develop universal standards, but it is not necessary to throw out the measurement baby with the bath water. It is possible to develop alternative forms of measurement and comparison that mitigate (if not resolve) the tensions outlined above and which allow for more democratic, less commodified and more explicitly *public* forms of service provision which can be used to compare and learn across place and time.

This will not be an easy political or technical task given the authority and resources of mainstream benchmarking organizations and their vested interests in existing models. So too will many public service providers resist change, particularly those that have sunk resources and political capital into current benchmarking frameworks. Even managers and policy makers who share the concerns raised in this paper may find it difficult to change directions given the inertia of existing benchmarking systems and the resources required to shift analytical and operational gears. It has taken decades of intense funding, lobbying and institutional support from major international organizations to get benchmarking systems to where they are today. It may take equally long to change them.

Nor will it be easy to sell the idea of revised benchmarking to those who are inherently sceptical of standardized, centralized evaluations (most notably those pushing for more autonomous forms of service provision). But a rejection of performance comparison vacates the possibility of more progressive and flexible notions of universality. Without some commonly agreed upon performance criteria how are we to know if a public service is "successful" and whether it should be celebrated? How do we articulate concrete demands for improvements on "equity" for example? How can we share "good" practice across different places? Rejecting benchmarking

altogether risks leaving this powerful organizational and discursive tool in the hands of those who may, intentionally or not, seek to advance the commercialization of public services.

These questions do not relieve us of the deep-seated philosophical and practical tensions between the universal and the particular, but they do force us to ask whether it is necessary to develop standardized criteria for public services. On this point I take my cue from Harvey (2000, p. 94) who notes, with reference to analogous debates around the development of standardized codes for international human rights, that "To turn our backs on such universals at this stage in our history … is to turn our backs on all manner of prospects for political action." In other words, rejecting universal standards for public services as important as water, health care, education and transportation runs the risk of not developing effective, concerted actions to deal with the hundreds of millions of people around the world with inadequate access to these basic amenities, possibly entrenching inequalities.

These debates also raise questions as to *whose* universalisms or particularisms matter. On this point I take my cue from Timmermans and Berg (1997, p. 275) who – writing about universalized medical protocols – argue that standardization efforts need not require a monolithic approach. Rather, universality should be seen as a "distributed activity" leading to "local universality …. an ambiguous and precarious status [that] emerges from localized processes of negotiations and pre-existing institutional, infrastructural, and material relations". In other words, universality can be a used in "non-transcendental" ways, "no longer implying a rupture with the "local" but transforming and emerging in and through it".

My proposals for new forms of performance evaluation are therefore driven by the need for standardized measurement principles that are supple enough to encourage local interpretation of "diverse goals, such as equity, stabilization, and social and environmental sustainability" (Lefeber & Vietorisz, 2007, pp. 139–140). The objective is to construct a dialectical bridge between these universalisms and particularisms, while recognizing that generalizations are inherently fraught with cultural and political tensions that disallow easy comparisons and may be irreconcilable at times. In this respect, the proposal for flexible forms of benchmarking in this chapter are consistent with, and complimentary to, the proposal for flexible notions of *publics* and public services outlined in Chapters 4 and 5.

This is not to suggest that mainstream benchmarking has ignored the tension between universal norms and local realities. Much of the mainstream benchmarking literature is at pains to highlight the need for "flexibility" and "local difference" (Baietti & Ginneken, 2006; Cabrera, 2008; Corton & Berg, 2009; Crotty 2004). The problem with these efforts, as I have argued above, is that they tend to be couched in an efficiency calculus that overwhelms all other factors, marginalizing questions of equity and promoting commercialization.

Alternative evaluation methods can never resolve these tensions of universality and particularity. Nevertheless, it is possible to be more explicit

about social, political and economic variability than mainstream benchmarking has been, particularly when it comes to market-oriented operating principles. Recognizing radically different cultural interpretations of services such as housing and waste management while prioritizing equity is possible but will require much more flexible systems of evaluation than those offered by current benchmarking models.

So too must alternative forms of benchmarking be more inclusive, "promoting participatory governance mechanisms which, when coupled with transparency, empower citizens when it comes to decisions on the management" of public services (APE, 2014, 2). Benchmarking processes and reporting will also need to take better account of technical issues such as literacy, numeracy and analytical skills as well as barriers to participation along socio-economic lines.

In Search of an Alternative

So where to begin? My proposal is to start with what already exists. There are many useful – and I would argue essential – indicators already in place in current benchmarking models that should be retained. In water services, for example, there are important metrics developed for water quality, infrastructure repairs, frequency of emergency breakdowns and a host of other relatively objective criteria which can be easily and usefully reproduced in new evaluation models.

What needs to change is (1) the ways in which we analyze some of these existing criteria, and (2) the addition of new indicators that better promote equity and explicitly promote a service's *public* nature. An example of the former is unaccounted-for water. Instead of simply tracking the number of leaky pipes and levels of non-payment, performance indicators should ask whether water services are affordable for low-income households and whether adequate investments are being made in bulk infrastructure in low-income neighbourhoods to prevent breakages. A more nuanced understanding of unaccounted-for water metrics could reduce harmful water cutoffs, improve the progressivity of tariffs and promote a more spatially equitable and transparent pattern of infrastructure investment.

On the second point, brand new indicators could be added to benchmarking systems, such as evaluations of worker health and safety, the scope of participatory decision-making mechanisms, and opportunities for female employees to move up the skills ladder. Such indicators could shed light on gender barriers to career progress in the public sector, provide insights into opportunities for improved community engagement, and improve the quality of work for labourers – important criteria that are virtually ignored in mainstream benchmarking systems.

The aim of these additional indicators is to evaluate the extent to which public service operators are offering change from marketized forms of public services, without which it is difficult (if not impossible)

to determine whether sufficient reforms have been made and whether a particular public service should be celebrated or not. As discussed in Chapter 5, the goal of restructuring public services is to shift our frameworks beyond the confines of a bourgeois public sphere to move towards less commodified and more equity-oriented service delivery. Non-marketized performance indicators are a concrete step towards developing objective yet flexible forms of evaluation that can help shape and entrench progressive policy directions.

It may also be wise to consider a much-reduced number of indicators than current benchmarking systems employ – already a topic of debate in the benchmarking literature. Van der Steen (2011, p. 33), for example, suggests that the number of performance indicators in the water sector should be in the 15-30 range. The City Blueprints for Water initiative uses 24 indicators: "a method that is practical, relatively simple, transparent, easy to communicate and understandable for decision-makers and the public in general", taking about a week to conduct (van Leeuwen et al., 2012, p. 2180).

Pidd (2009) promotes the principle of "model simple, think complicated" (drawing on Little, 1970), arguing that benchmarking systems should be easy to understand, robust enough to allow counterintuitive results to emerge, simple to manage, adaptive to different situations, and easy to communicate. The challenge is to find a balance between the complex reality of water systems and the need for simplification in ways that "helps focus people's minds" (Pidd, 2012, pp. 75–6).

There are, of course, good reasons why performance evaluation has grown in complexity over the years, and my proposals for more progressive metrics would only compound this. However, *more* measurement does not necessarily mean *better* measurement. Just how many indicators are optimal is not clear, but the enormous volume of measurement indicators promoted by the ISO, and the byzantine statistical analyses that accompany them, are too complex and too opaque for meaningful public debate, overwhelming even the best-resourced of public service providers. An alternative benchmarking framework may therefore benefit from a more streamlined and popularly accessible approach.

I would, however, warn against false simplicity. AquaRating (2013, p. 5), for example, has developed a benchmarking model that provides a single evaluation score (a number between 1 and 100), but the calculations required to arrive at this figure are as complex as ever: "eight assessment areas, 27 assessment sub-areas, 113 assessment elements, 61 indicators, 99 variables, 52 practices groups, [and] 348 individual practices." This "simple" benchmarking system still requires experts and enormous time and resources while its results remain impenetrable to the average front-line worker and water user – presumably one of the reasons that they charge between US$25,000 and US$65,000 for an audit (AquaRating, 2014b, p. 6).

A Spider Web Evaluation

The alternative evaluation framework I am proposing here (in very broad and preliminary terms) attempts to find a middle ground between the need for simplicity and the reality of complexity on the ground, while at the same time employing new and old evaluation methods. My intent is not to suggest a final, polished product, but rather to concretely demonstrate the possibility of a more user-friendly form of data collection and analysis that also visually represents results in ways that promote public engagement and accessibility. In doing so I draw on the existing practice of clustering benchmarking categories (such as EBC's performance areas of "water quality, reliability, sustainability, service, and finance/efficiency" (EBC,2014, 7)) and efforts by groups such as City Blueprints for Water to simplify and pictorialize their benchmarking system (van Leeuwen et al., 2012, p. 2180). The latter employs "spider diagrams", which I have adopted here, although other visual representations could be equally effective.

Figure 6.1 is an example of such a benchmarking representation, once again using water services as an illustrative example, comparing two hypothetical public water operators. It employs nine overarching performance categories, each with a series of easily understandable indicators. Visually, the longer the shaded area on each strand the better the service for that criterion, and the larger the overall shading on the web the better the overall performance of the operator.

Table 6.1 outlines the proposed indicators and metrics that could make up these valuations. Once again, these are not necessarily the best or the

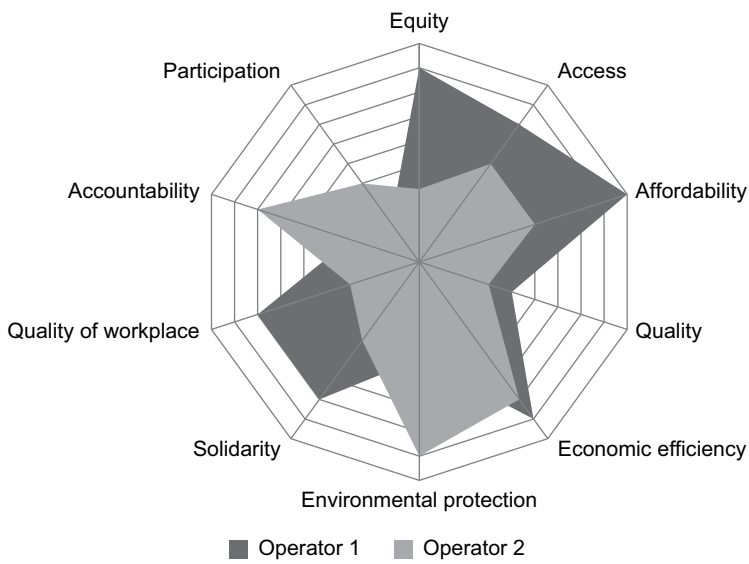

Figure 6.1 Spider Diagram Comparing Public Water Operators

Table 6.1 Normative Criteria for Evaluating Public Service Providers

Normative category	*Criteria*	*Definition*	*Examples of possible sub-criteria questions*	*Examples of possible measurement indicators*
Universality	Access	Physical availability of the service at a convenient distance from user's dwelling	• Rural/urban divide? • Sufficient quantity? • Culturally acceptable service?	• Proportion of population with adequate access • Time-distance to service location • Hours/day that service is available
	Affordability	Prices that ensure economic accessibility for all	• Are poorer households disproportionately burdened? • Are programs in place for cross-subsidy pricing? • Is affordability a legal obligation?	• Cost as percentage of household income • Disconnection rates • Levels of subsidization by region
	Quality	Reliable, satisfactory services that create positive relations with end users	• Safe for all users? • Responsive to user needs? • Ongoing improvement mechanisms in place?	• Primary health outcomes • Level of service interruptions • Complaints by region
	Equity	Equality of opportunity to access quality services for all	• Equitable quantity of service across user groups? • Equitable quality of service across user groups? • Is equity formalized, legalized or institutionalized?	• Budget allocations by region • Levels of access by socially disadvantaged groups • Per capita consumption by region

(Continued)

Table 6.1 Normative Criteria for Evaluating Public Service Providers *(Continued)*

Normative category	*Criteria*	*Definition*	*Examples of possible sub-criteria questions*	*Examples of possible measurement indicators*
Sustainability	Efficiency	Cost-effective use of resources to meet service mandates	• Are current infrastructure investments helping to meet the social goals of the service? • Is the capital intensity of investments appropriate? • Do short-term cost reductions undermine long-term efficiency gains?	• Financing as a proportion of overall operating costs • Cost per unit of service delivered by region • Employee turnover rates
	Environmental protection	Meeting current service mandates without compromising future resource needs or undermining current environmental sustainability	• Are programs in place to reduce demand on natural resources? • Does the service provider respect different cultural understandings of resources? • Are climate change mitigation plans in place?	• Levels of renewable energy use • Quality of wastewater treatment • Rates of respiratory infection
	Solidarity	Cohesion among various producer and user groups and across sectors that builds economic, social and political commitment to a public service model	• Does the model help to build a stronger "public ethos" around services? • Is the service contributing to improvements in other sectors and at other levels of service delivery? • Does the service model explicitly oppose privatization and commercialization, with sufficient political support?	• Formal cooperation agreements between different levels of government and sectors • Measurements of inter-sectoral impacts (e.g. sanitation extension reducing diarrheal burden)? • Legal mechanisms to prevent privatization

(Continued)

Table 6.1 Normative Criteria for Evaluating Public Service Providers *(Continued)*

Normative category	*Criteria*	*Definition*	*Examples of possible sub-criteria questions*	*Examples of possible measurement indicators*
Governance	Accountability	Obligation to account for activities, accept responsibility for them, and to disclose the results in a transparent manner, readily available to the public, and understandable.	• Are there clear operational mandates and policy positions? • Are there transparent capital and operating budgets? • Are mechanisms of accountability available at appropriate scales (local, national, regional)?	• Transparency of hiring processes • Access to mechanisms of accountability by region • percent of documentation openly available and verifiable, in suitable languages and formats for all users
	Participation	Citizen involvement in policy making and implementation of service delivery	• Is participation at appropriate scales and sufficiently representative? • Are there adequate resources for participation by a diverse range of society (transportation, time off work, etc.)? • Is participation conducted in culturally appropriate ways?	• Number of people participating in formalized mechanisms of participation • Number of different processes of participation open to participation (policy making, budget decisions, etc.) • Availability of participation by region
	Quality of Workplace	A place of work that provides a safe environment, trust between employees and management, fairness, and a sensible workload that contributes to quality service delivery	• Are there adequate numbers of workers to ensure service quality? • Are there mechanisms for workers/unions to participate in the operation, management or policy-making of the service? • Are there good feedback loops between front-line workers, managers and end-users of the service?	• Pay equity (job type, gender, race, ethnicity, etc.) • Availability of health and safety equipment • Access to training opportunities

only benchmarking criteria – or the correct number of categories – but they serve to demonstrate what an alternative system might look like. Overall, the aim of these indicators is to prioritize what a colleague and I have previously described as cornerstones of a public service – universality, sustainability and democratic governance (McDonald & Ruiters, 2012b). Financial resources are important, but they should not overshadow other priorities, and should not be confined to a single sector. Water services, for example, are intricately linked to health and well-being across a range of services, requiring a multi-sectoral, multi-scalar and even multi-jurisdictional perspective.

The criteria and indicators proposed in Table 6.1 are suggestive of the kinds of evaluation questions and techniques that are aimed at equity and which go beyond the confines of water to look more broadly at the public goods that are produced (or not) from the provision of water services. The intent is to make more "public" the otherwise narrow institutional and commercialized evaluation criteria found in existing benchmarking models. I have also introduced a category for solidarity, defined here as cohesion among various producer and user groups and across sectors that build economic, social and political commitment to a public service model. Evaluation of this category could include such questions as: Does the model help to build a stronger "public ethos" around services? Is the service contributing to improvements in other sectors and at other levels of service delivery?

The questions outlined in Table 6.1 are not the only or the most exhaustive ones that could be asked about the "publicness" of a water (or other service) operator. They do, however, serve as a counterpoint to the commercial bias of existing benchmarking models. They also make explicit the need to evaluate a service provider's ability to address broad public good outcomes. In this respect, it would be interesting to see how *private* water operators perform using this *public* scorecard – an inverse of current benchmarking realities!

These proposed criteria are not without their limitations or potential abuses of course. It is possible for assessors to impose alternative performance indicators in crudely universalistic ways, with little interest in local variation or cultural diversity. These new criteria would, however, force public service operators to grapple more explicitly with provocative questions of inequality, accountability and sustainability in ways that could force more meaningful and contextualized change. A revised benchmarking system such as this could also be used to promote better engagement with service users, contributing to a spirit of local ownership of performance indicators that is largely absent from the highly centralized, top-down and expert-driven benchmarking models in use today.

Conclusion

As a technical exercise, benchmarking may seem objective, but in practice the selection of measurement criteria used for public services, and the ways in which they are employed, are anything but. Dominant forms

of benchmarking valorize the commodity value of public services, exclude the public from participation, and marginalize alternative worldviews. Deliberately or not, a handful of powerful international actors have designed monopoly-like forms of performance evaluation that emerged from private sector practice and which are growing in influence and scope in public services.

If we are to change the ways in which we operationalize public services and move away from commercialization it is essential to change the ways in which they are evaluated. If new public operators are appraised by the same marketized criteria as their predecessors it will be impossible to change outcomes on the ground as old performance evaluation systems reward and reproduce themselves.

Developing substitutes will not be straightforward, particularly if they are seen as a challenge to market norms. Nor will it be easy to convince public managers and policy makers to adopt new frameworks given the time and resources required, even if they are ideologically inclined to do so. The practical and political reality of building alternative benchmarking frameworks is an admittedly daunting one, but it is a challenge that cannot be ignored.

Notes

1. http://www.iwapublishing.com/template.cfm?name=m2477&utm_source=IWA+Publishing+Mailing+List&utm_campaign=e3de7ca958-GND_9_September_2014&utm_medium=email&utm_term=0_49a7734030-e3de7ca958-90138637 accessed September 9 2014.
2. See www.waterbenchmark.org/news/EBCcontributestoMDG-proje.html
3. IBNET – http://www.ib-net.org/en/texts.php?folder_id=78, accessed August 2 2014.
4. See www.iwabenchmarking.com/site/documents/
5. See http://www.ib-net.org/en/texts.php?folder_id=91&mat_id=72&L=0&S=0&ss=0, accessed August 7 2014.
6. http://www.ib-net.org/en/texts.php?folder_id=78 accessed August 5 2014.
7. http://www.ib-net.org/en/texts.php?folder_id=78
8. http://www.iso.org/iso/home/standards_development/standardsdevelopment_gettinginvolved.htm and http://www.iso.org/iso/home/about.htm