The Art and Science of Benchmarking

by

Sanford Berg
Director, Public Utility Research Center
www.purc.org

Abstract: *Infrastructure is in need of rehabilitation in many developed countries. In emerging countries, water has come to symbolize the huge gaps between promise and performance. The sad truth is that the political economy of water is such that those who make tough decisions will not receive credit during their terms in office. Nevertheless, benchmarking represents an important tool for documenting past performance, establishing baselines for gauging improvements, and making comparisons across service providers. There are at least six audiences for yardstick comparisons. First, benchmarking specialists produce and critique studies that utilize various methodologies. Second, the press filters and highlights reports, using "sound bites," executive summaries, and interviews. Third, the general public tries to understand the implications of rankings for their own evaluation of sector performance. Fourth, the regulator reviews studies and creates performance incentives to achieve policy objectives. Fifth, national policymakers (elected representatives and appointed officials) react to and utilize technical studies in setting priorities and interacting with international organizations. Sixth, water utility managers are sensitive to comparisons as they have much to lose (and something to gain) when information is made public. Although each group has different needs, relative and absolute rankings can become catalysts for better stewardship of water and other resources.*

NOTE: This paper was originally prepared for presentation at the conference on Global Developments in Water Industry Performance Benchmarking held September 29, 2003, in Perth, Australia.

A series of international meetings has drawn attention to serious global problems of water and wastewater. In many developed countries, infrastructure is in need of rehabilitation. In emerging countries, water has come to symbolize the huge gaps between promise and performance. Lives are lost each day when serious reforms are delayed and investments deferred. The sad truth is that the political economy of water is such that those who make tough decisions will not receive credit during their terms in office. Decisions about financing and operating water systems involve time horizons that are much longer than the number of months before the next election.

To be successful, the international debate must move from rhetoric to reality. Benchmarking is about reality—relative performance through yardstick comparisons. It is one of the most important tools available for documenting past performance, establishing baselines for gauging improvements, and making comparisons across service providers. By definition, half of any set of comparable utilities will be below average (the median). Thus, the results of a significant benchmarking study can be very disruptive. Some managers will feel threatened by comparisons since reality can hurt. Other managers should reap rewards as those receiving service begin to appreciate the excellent job being done to hold down costs, maintain financial sustainability, and improve service quality.

The validity of performance scores (in an absolute sense) and benchmarking comparisons depends on whether the studies utilize best practice. Currently, there are significant international initiatives to collect comparative data. Even more important than data collection is how the regulator, national development banks, and international funding agencies utilize scorecards to reward good performance and put pressure on poorly performing utilities (private and public). We can collect vast quantities of technical data, but only when patterns are carefully analyzed do the scorecards become convincing. However, not only must the studies be technically sound, they must be carefully communicated so the public can place pressure on those who are not being good stewards of water resources. Also, regulators generally utilize the results when they set prices and/or expansion targets for utilities: incentives need to reflect reality. In addition, scorecards must be used by multilateral funding agencies to screen loan candidates. If poorly performing utilities are rewarded with financial subsidies while those engaged in cost-containment are merely congratulated, the resulting perverse incentives will weaken average performance and result in continuance of wasted resources. Finally, a benchmarking study that "names and shames" a water utility has implications for its managers and perhaps for their job tenure. So it is important that the analysis of the data used in the scorecard is sound and that the information be used in a way that is defensible.

Unless the reality revealed by yardstick comparisons has serious financial consequences for water utilities, there will be a continuation of business as usual. In developed nations, city water systems will continue to deteriorate until massive national subsidies get spread across the board to politically favored cities and to water utilities that delayed making tough decisions. In emerging markets, the poor performance syndrome will continue: "consumers will pretend to pay and utilities will pretend to supply good service." Private

capital will be unavailable for network expansion because the underlying financial conditions of water/wastewater utilities will be based on promises, not on performance.

The message is simple: without the political will, benchmarking studies become just another set of reports that have little impact on industry performance. So while improving data collection and analysis are important, we might do well to consider how a small set of indicators could be utilized more effectively.

There are at least six audiences for yardstick comparisons: benchmarking specialists, the press, the general public, the regulator, policymakers (elected representatives and appointed officials), and water utility managers. Although each group has different needs, all can use the relative and absolute rankings as catalysts for better stewardship of water and other resources. *Relative* rankings allow the different audiences to compare the performance of utilities in comparable situations. Here, the key problem is how to select firms that are truly similar to one another. Alternatively, how can the rankings reflect the different conditions managers face. We want the relative ranking to reflect managerial decisions rather than the unique characteristics of service territories beyond managers' control, including topography, hydrology, and density. In addition, history matters: current managers inherited utility systems that reflected a set of political and economic (including regulatory) decisions made by others. Thus, performance improvements over time also need to be taken into consideration. In addition, *absolute* comparisons are necessary, since the weakest performer in one group might have much better performance than the best firms in another group of comparable firms (say, those in another country at a similar stage of development). Comparisons are valid so long as the results do indeed tell us whether particular firms are performing far below potential. At issue is how to define "potential performance."

Benchmarking Specialists Produce and Critique Studies

Currently, quantitative comparisons utilize four methodologies: sophisticated average costs and productivity indicators, linear regression analysis, Data Envelopment Analysis (DEA—using non-parametric cost frontiers), and the economic/engineering analyses ("model company" or "virtual firm" method). Data reliability and comparability affect the validity of all four approaches.

Rankings can be manipulated by choice of variables, model specification, sample size, time frame, and treatment of outliers. Because the stakes are high, affected parties have an interest in the relative and absolute performance evaluations prepared by analysts. Studies can be controversial. Even if everyone agrees on the scores, the implications for setting prices or targets are not straightforward. Implementing policy means making judgment calls. Technical experts may not possess the skills required to negotiate agreements regarding how rankings are to be used. Ultimately, those preparing studies must recognize that they are writing for multiple audiences with a variety of needs.

The Press Filters and Highlights Technical Studies

Let us assume that best practice has yielded a defensible and convincing set of performance indicators for water utilities in a country. Unfortunately, technical reports are not amenable to sound bites. Nevertheless, most newspaper reporters and television journalists seek the clear message that emerges from a benchmarking study. Yet, some reporters seek sensational factoids that support their own ideological predilections while others lack the expertise to interpret technical studies.

We all know that quantification does not always shed light on a topic. Rather, a thoughtful comparison of utility performance is likely to emphasize the complexities inherent in the numbers. Benchmarking is not a discipline like physics; it is a mix of economics, engineering, statistics, finance, and many other fields. A thoughtful distillation of a long report will note the tentative nature of any conclusions. Yet, citizens are going to use a League Table to evaluate water utility managers.

I view benchmarking as a long-range activity. The first rankings are bound to contain errors since any new data collection effort requires a new set of activities for utilities and the development of procedures for auditing numbers that are supplied. One can argue that well-managed firms should already have records with basic data. However, consistency of definitions and processing rules both will require some fine-tuning when nations (and water regulators) become serious about benchmarking.

In *How to Lie with Statistics*, Darrell Huff demonstrates how statistical patterns are affected by sample selection and shows how averages (mean, mode, and median) must be carefully interpreted. Perhaps the most humorous elements of the book relate to how charts can present data in a distorted manner.[1] Clearly, the media is susceptible to both deliberate and unknowing bias. What are the implications for those preparing and distributing yardstick comparisons? The problem of communicating complexity places a special burden on those preparing benchmarking reports. The scores need to reflect reality and to have serious implications for the best and weakest performers. The executive summary must focus on the implications of the study, while recognizing the limitations in the analysis.

The Public Needs to be Educated about Complexity and Performance

The media are not the sole filter for studies. Opinion leaders can be reached through other channels. However, public education is likely to be heavily influenced by what people read and hear in the news. This observation suggests that long before releasing a benchmark comparison, the responsible agency should be engaging in an information dissemination campaign—informing political leaders and the press about the purpose of the forthcoming report. The project might release a set of studies over a longer time

[1] When we enter the world of empirical studies, we know that ". . . if you torture the data enough, it will confess." Playing with model specification and data "outliers" enables the skilled (but dishonest) statistician to prove what he or she set out to prove. See *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists* by Joel Best.

period, laying the foundation for the League Tables and other (potentially) sensational aspects of the study when the results are finalized. The news stories that accompany the series of studies could then focus on specific indicators regarding water resource use, operational performance, quality of service, and financial sustainability.

Another set of channels are the non-governmental organizations (NGOs) that monitor and participate in water sector regulatory decisions. NGOs and formal Citizen Advisory Committees (that might be established by regulatory commissions) provide opportunities for input and feedback for citizens. At the local level, such groups are particularly valuable forums for educating citizens and for learning about their needs and concerns. Thus, a benchmarking report has multiple audiences with different interests and needs which suggests that several formats might be utilized to bring the results to the citizenry.

The agency responsible for the study must choose between releasing one large study, where the news stories are likely to focus on the most sensational elements (for headlines), or a series of studies, where the cumulative effect of the reports may be a informed citizenry. However, in the latter case, the impact may be diluted as the attention of policymakers and the press turns to other events. Ultimately, those paying for water and sewerage need both to appreciate the complexity of the issues and to understand the implications of the rankings for their own local water utility. The best strategy for releasing results is not clear.

Regulators Create Incentives and Affect Performance

Firms react to incentives, including those that emerge from the timing of rate reviews. Jamasb, Nillesen, and Pollitt (2003) use DEA to test for strategic behavior by electricity distributions companies. Some behavior involves presenting the firm's performance in a favorable light—by reclassifying some types of expenditures or smoothing some accounting numbers. This activity might transfer money from customers to the firm, but cost minimization is still encouraged. However, some strategic behavior can lead to inefficiencies; when these are detected, the regulator must revisit the formal incentives. Yardstick comparisons provide one framework for detecting inefficiencies, but gaming also holds the potential for influencing the performance scores of other firms.

Productivity advance and other measures of technical efficiency provide valuable information for policymakers. However, excessive simplicity can result in a distorted analysis. Sector performance also depends on efficient price signals, benefits from quality improvements, incorporation of environmental impacts into decisions, recognition of transition costs, and on meeting social obligations. These considerations are discussed below.

i. ***Pricing Efficiency:*** Setting price equal to marginal cost is a standard for allocative efficiency. When making actual sector evaluations, the analyst must incorporate both short-run and long-run cost considerations into the evaluation of the prices. Returns on investment cannot be ignored, since a short-run price signal that is efficient is not necessarily sustainable. Rate design can result in

marginal price tracking marginal cost, with fixed costs recovered via monthly customer charges or hookup fees. Thus, a complete benchmarking analysis should address pricing issues, and outline a framework for incorporating rate design into performance evaluation.[2]

ii. ***Service Quality*****:** Clearly, outputs of different quality need to be treated differently when evaluating performance. DEA is particularly useful in incorporating multiple outputs (including dimensions of quality) into comparisons. The policy issue is whether being "best" at delivering a level of service quality that is not valued highly should be allowed to generate a high score for the utility. An example of a service quality index is the Overall Performance Assessment used by Ofwat to compare 23 water and water and sewerage companies. That index includes eight measures. Ofwat assigns a score falling in the range of good, acceptable, or needs improvement for five quality dimensions related to water supply and customer service.

iii. ***Recognizing Transition Costs***: Over the past decade, most nations have targeted certain types of infrastructure for restructuring. Care must be taken when making international comparisons since restructuring is disruptive to management and can involve costly investments (such as dealing with historically underfunded pensions and insufficient maintenance in the past). Investments can include developing internal quality monitoring systems, creating modern information/billing systems by small suppliers, or staff training to help the firm meet new expectations. Whether aggregate data will be affected by associated expenditures depends on the extent of such programs.

iv. ***Environmental Impacts and Water Resource Management*****:** Other dimensions of performance that can be missed by simple input-output comparisons are externalities. A firm can cut back expenditures today and impose costs on future customers. The simple numbers in accounting ledgers will not capture the full consequences of deferred pipe maintenance or delayed replacement of pumping equipment. Benchmarking that ignores environmental impacts on groundwater, which affects future costs, provides inappropriate comparisons. Simple comparisons using aggregate numbers will be misleading unless environmental impacts are not included in the analysis.[3]

v. ***Social Obligations*****:** The suppliers of water/wastewater services in different jurisdictions are likely to have different service obligations. Some will serve both urban and rural populations, while others serve populations with a narrow range of population density. Service to low-income customers can involve issues of non-payment, cross-subsidies, and commercial losses. All three of these situations affect the financial health of the supplier. Thus, the production costs

[2] Chapters in Dinar (2000) underscore the importance of water price reform for sector performance.

[3] Saleth and Dinar (1999) explore how water law, water policy, water administration, and other institutional features affect water sector performance. From this perspective, simple performance scores are likely to miss key determinants of the sustainability of sector performance.

> associated with meeting different types of service obligations need to be incorporated in any efficiency comparisons. When cash flows are low in the presence of mandated service obligations, the firm may defer maintenance or experience significant delays in network expansion. These issues need to be addressed when conducting a performance benchmarking study.

What does a regulator do with a benchmarking study? Shuttleworth (2003, p. 50) is highly critical of regulatory use of yardstick comparisons produced by researchers:

> I suspect that attempts at benchmarking will only be abandoned when regulators realize that it doesn't help them. Even if one believes that a network's costs should be 20 percent lower than they are, it is a heroic assumption that such inefficiency (no matter what its level) will or can be eliminated over the course of a rate plan (no matter how its length is determined or what it is). In practice, such decisions raise important questions about what is a reasonable rate of cost reduction—which is where the discussion began. Attempts to impose benchmarking solutions without solving these fundamental problems must rely on arbitrary disallowances, or else command-and-control methods of reducing utility prices and costs. Either approach would destroy the incentive properties of regulation using a price cap formula.

Thus, benchmarking might not be a reliable means for establishing rates. However, studies can certainly point to potential problems, and indicate where more in-depth analysis is needed. A ranking or score is not a Rosetta Stone that automatically translates one set of data into pricing formulas.[4]

Policymakers React to and Utilize Studies

Two types of policymakers can influence the development of the water/wastewater industry. Domestic political leaders are confronted with performance indicators that either reflect well on managers (and the policy environment) or indicate that the system is not meeting performance expectations. Just as important, when making soft-loans (subsidies to wastewater treatment, especially), multilateral lending agencies can take performance indices into account.[5]

Solid data about infrastructure performance provides an antidote to government opportunism and the politicization of water pricing. Those making promises will, at some point, be held accountable if reality can be shown to deviate from past rhetoric. Information is like a cleanser—it makes visible the actual performance of public and privately owned water utilities. One might ask why benchmarking performance studies have not been significant tools in most countries over the past four decades. One answer

[4] Regional regulatory associations can provide cross-country comparisons that help identify areas for improvement. ADERASA in the Americas, AFUR in Africa, and SAFIR in South Asia promote useful networking among regulators and benchmarking specialists.

[5] For an excellent overview of the political economy of water and a number of case studies, see Shirley (2002).

is that people in political power did not really want to have the information. The resulting financial crisis experienced by many water utilities has led to increasing private sector participation and multilateral funding initiatives. These developments have been accompanied by calls for greater accountability to protect external investors. Independent regulatory commissions have been created to provide tools and a time horizon that are more appropriate for addressing the political economy of infrastructure. When industry oversight responsibility is assigned to well-trained regulatory professionals, the resulting benchmarking studies can truly make a difference in terms of public attitudes and perceptions.

The tragedy is that aid programs and multilateral organizations were slow to utilize benchmarking in the 1970s and 1980s for a rational triage approach to funding infrastructure investments. The resource misallocation and corruption that accompanied much of the Cold War-inspired aid is a sad reflection on the priorities of the developed nations. A strong argument can be made that the weakest performing systems should not be given aid, and that high-performing systems should be rewarded. This conclusion seems heartless, yet unless tough funding decisions are made, incentives are quite perverse. Those nations and utility systems making the tough decisions deserve to be rewarded. In circumstances where lives are at risk (the poorest of the poor), emergency measures are certainly warranted. However, one can argue that this aid should not involve significant untargeted outlays, but only short-term (targeted) funds. First, civil society must develop the institutions, political consensus, and the critical mass of expertise required to create sustainable organizations. Capital-intensive projects might be least-cost in principle, but they are likely to result in delays and waste that dissipates most of the potential benefits from such aid. I wish I knew how to manage this terrible problem: for now, I can only point out that the historical record is not one that gives me much confidence.

One possible application of benchmarking involves projects with "hard" budgets. might International lending institutions could buy bonds issued by water utilities (where eligibility depends on past performance and a commitment to benchmarking). Let me cite a simple example. Assume that the "true" cost of capital is 12%. Form a payment schedule that involves a direct subsidy. The terms are clear: the utility (through rate-payers) will start paying 5% the first year, 6% the next, etc. If the funds are used for productive projects (and consumers start paying the true cost of service), by year eight, the project is sustainable.

Annual Interest Payments

Rate-payers	5	6	7	8	9	10	11	12
Subsidy	7	6	5	4	3	2	1	0

If this is a 20-year bond, it is now seasoned. Private investors can evaluate the track record of the water utility. Benchmarking information is available on financial and other performance indicators. If the perceived commercial risk associated with the bond has fallen, the "true" cost of capital might now be 8%. The lending institution can sell the bond for a higher price than the principal, since the bond has appreciated in value. The

proceeds can be used to cover some of the explicit subsidies in the initial two-part support system. This example is only meant to illustrate the potential benefit of making aid funds conditional on benchmarking outcomes and on meeting hard (not soft) financial targets.

Utility Managers Have Much to Lose when Information Is Made Public

It is extremely difficult for an outsider to evaluate the performance of managers. Inadequate reports and the selective presentation of information mean that only insiders know whether the organization is managed well or poorly. Benchmarking reduces the extent of this information asymmetry. For this reason alone, utility managers might delay or block serious benchmarking initiatives. However, many utilities are well-managed, and it is in the interest of some executives to promote comparisons that enhance *their* careers. Nevertheless, yardstick comparisons can be used to put pressure on all managers to improve performance. Such a threat is likely to dim interest in making data available that could serve as a catalyst for change.

In addition, information is costly to produce unless it is a by-product of on-going management processes. Berg and Corton (2002) identify wide range of information required for managers and financial institutions need to evaluate performance: technical and operational data, commercial and financial information, staffing numbers, and environmental impact reports. One can ask why some historical data are so inadequate. One answer is that some systems have been poorly managed—where poor data provide evidence of lax management. However, oversight institutions also bear some responsibility for poor data. Government ministries and international funding agencies are happy to cut ribbons at grand openings, but the difficult day-to-day operations of water systems do not lend themselves to public ceremonies. If the promises of politicians are never confronted with the reality of performance outcomes, then one can understand the weak performance (low-level equilibrium) identified in the Savedoff and Spiller (1999) volume on Latin American water systems. Thus, serious benchmarking may not be in the interest of those agencies that are supposed to be monitoring the sector.

Concluding Observations

The first (Berg, 2001) in a series of six articles written for *Water21*, the journal of the International Water Association, in 2001-02 began with the following statement:

> There is no simple "solution" to the water problems faced in developing economies. However, problems can be managed more effectively if some oversight responsibility is assigned to an independent regulatory commission. This article highlights a fundamental issue regarding water sector reform facing nations today: how to constrain political opportunism. All stakeholders--equipment suppliers, multi-lateral agencies, water systems operators, customers, and government ministries--have an interest in improving the investment climate so nations can move forward to build and operate the water systems needed for the 21st century.

No problem can be managed in the abstract. Benchmarking is no different. The art and science of yardstick comparisons requires technical expertise and experience, and benchmarking is one of tools that can make a difference for the future.

Bibliography

Alegre, Helena, Wolfram Hirnir, Jamie Melo Baptista and Renato Parena, (2000). *Performance Indicators for Water Supply Services,* IWA Publishing, xiii-146.

Anderson, Timothy R., Rafael Borja, Ivan Patricio Hernandez, Fabricio Tobar, Lioino Setiowijoso, (2003). "Extending Productivity Research Frontiers: DEA Resources of Datasets and Errata," *Journal of Productivity Analysis, 19, 271-275.* http://www.etm.pdx.edu/dea/dataset/

Berg, Sanford (2001). "Consumers Pretend to Pay and Utilities Pretend to Supply Good Service: Breaking the Poor Performance Syndrome," *Water21,* October, 64-65.

Berg, Sanford and Maria Luisa Corton (2002). "Infrastructure Management: Applications to Latin America," in *Private Initiatives in Infrastructure: Priorities, Incentives, and Performance,* Sanford Berg, Michael G. Pollitt, and Masstsugu Tsuji eds. Edward Elgar Publishing, 189-202.

Corton, Maria Luisa (2003). "Benchmarking in the Latin American Water Sector: the case of Peru," *Utilities Policy,*11, 133-142.

Dinar, Ariel ed. (2000). *The Political Economy of Water Pricing Reforms,* World Bank, Oxford University Press, x-405.

Jamasb, Tooraj, Paul Nillesen and Michael Pollitt, (2003). "Strategic Behavior under Regulation Benchmarking," DAE Working Paper 0312, Department of Applied Economics, University of Cambridge, January, 1-28.

Saleth R. Maria and Ariel Dinar (1999). *Evaluating Water Institutions and Water Sector Performance,* World Bank Technical Paper No. 447, xi-93.

Savedoff, William and Pablo Spiller, eds. (1999). *Spilled Water: Institutional Commitment in the Provision of Water Services,* Washington, D.C.: Inter-American Development Bank. 1-248.

Shirley, Mary, ed., (2002). *Thirsting for Efficiency: the Economics and Politics of Urban Water System Reform,* The World Bank: Pergamon. xxi-376.

Shuttleworth, Graham (2003). "Firm-Specific Productive Efficiency: A Response," *The Electricity Journal,* April, 42-50.

Zhu, Joe. (2003) *Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver*, Boston: Kluwer Academic Publishers, xxiii-297.

.