Multistakeholder Evaluation of Condominial Sewer Services

Earthea Nance, *Texas Southern University*

Multistakeholder Evaluation of Condominial Sewer Services

Earthea B. Nance
Virginia Polytechnic Institute and State University

Abstract: A multistakeholder evaluation procedure is presented to address the many challenges in evaluating the performance of condominial sewer projects in Brazil. Condominial sewerage is a promising appropriate technology that is coproduced by users and public agencies, but little is known about project performance. This article shows that multistakeholder evaluation is more appropriate to the research setting than evaluations based on a single stakeholder's perspective, is more reliable than the use of expert judgments alone, and provides more information than recommended sewer performance indicators. The author argues that the perspectives of multiple stakeholders should be included in project evaluations. The multistakeholder approach presented is a workable solution to the condominial sewer evaluation problem.

Keywords: *project evaluation; multiple stakeholders; performance indicators; condominial sewers*

The Evaluation Problem

This article is about understanding how a system performs by uncovering how it is used, drawing on condominial sewers as an illustration. I describe an approach to evaluation that better matches the service provision process and that expands the criteria of merit for determining the value of projects such as these. In so doing, the multistakeholder evaluation approach I propose represents an advance over past practices that favored only expert stakeholders' perspectives and technical criteria. The examples presented herein, although of particular interest to evaluators of international development projects, offer larger lessons for evaluators working in different substantive areas and in different settings, namely, that in complex evaluation settings with heterogeneous views toward evaluands, it is important to include multiple perspectives and to evaluate programs by criteria that reflect these different stances.

Condominial sewers are relatively new to Brazil's sanitation sector. Developed in the early 1980s as a feasible alternative to conventional sewers, condominial sewers are physically characterized by inexpensive materials, shallow excavations, small diameters, and modest slopes. The systems are located in backyards, front yards, and sidewalks to serve blocks of homes collectively. Collective service reduces the overall length of sewer pipe needed and facilitates sewer service to homes in unplanned areas that lack distinct blocks or paved streets. The alternative design results in significant capital cost savings compared with conventional sewers.

Earthea B. Nance, Virginia Polytechnic Institute and State University, School of Public and International Affairs, 211 Architecture Annex (0113), Blacksburg, VA 24061; e-mail: ebnance@vt.edu.

American Journal of Evaluation, Vol. 26 No. 4, December 2005 480-500
DOI: 10.1177/1098214005281324

Condominial sewers are also characterized by a participatory implementation process, in which beneficiaries are actively involved in the planning, design, construction, and maintenance of the systems.

Despite the great potential of condominial sewers to expand access to basic sewer service in Brazil and other developing countries (World Bank, 1994), our understanding of actual performance outcomes is limited. Watson (1995) described performance at the city level, and Ostrom (1996) theorized about the conditions under which condominial sewers should succeed. There is little, if any, evaluation research on the quality of condominial sewer project performance.

Evaluation scholars point out that evaluation, which is a process for assigning value (Mohr, 1999), is important for several reasons. Evaluation allows stakeholders and outside evaluators to learn from practice, to build the capacity of residents involved in coproducing urban services, and to enhance accountability among service delivery organizations and residents (Chelimsky & Shadish, 1997; Fetterman, 1997). Evaluation allows stakeholders and outside evaluators to check the assumptions under which a project was implemented (Birckmayer & Weiss, 2000) and helps stakeholders assign value to various project inputs on the basis of knowledge of the outcomes (Mohr, 1999). Conclusions about what works in the implementation process and about condominial sewer technology depend on how outcomes are interpreted.

If evaluation is important, then why is there so little information about the performance of condominial sewer projects? In 2002, the U.S. Environmental Protection Agency (EPA) reported that sewer system evaluation in general is not commonly done, that standard sewer performance indicators do not exist, and that methods for evaluation are poorly developed. Formal sewer evaluation approaches developed by the EPA (1977, 1991) and the American Society of Civil Engineers (1994), which involve the gathering of instrument data and the development of hydraulic models, are rarely conducted in developing countries.

There have been several attempts to develop standard sets of performance indicators in the sanitation sector (e.g., American Society of Civil Engineers, 1998; National Council on Public Works, 1988; U.K. Office of Water, 1998; World Bank, 2004). Table 1 presents a sample of recommended sewer performance indicators from the engineering literature. These indicators reveal a primary concern with service output and investment efficiency and represent the perspective of a single stakeholder group (i.e., the sewer officials who conduct and use the evaluation). The perspectives and concerns of other stakeholders (e.g., neighborhood residents, maintenance workers) are either minimally present or altogether absent. Recommended indicators such as those in Table 1 are typically not presented within a conceptual framework. Such a framework would help sort out the dimensions of performance and would suggest the kinds of observations and measurements needed to represent the different aspects of performance. Without a conceptual framework, it is not clear which set of indicators should be used in a given setting, how the indicators relate to one another, or why certain categories of indicators have been excluded. Many of the indicators in Table 1 provide no information about how users value service outputs, and they provide little, if any, information about the quality of service outcomes, particularly with regard to health and environmental improvements. On the basis of the indicators in Table 1, the accepted meaning of sewer performance is limited to technical and financial dimensions.

Public administration literature offers an alternative approach to performance evaluation. That literature encourages the mixing of objective and subjective information about performance (Brudney & England, 1982; Kelly & Swindell, 2002a) and the inclusion of multiple stakeholders' perspectives on the basis of the view that relying on a single indicator is flawed (Ostrom, 1973). Some studies have shown correlations between subjective and objective outcome measures (Parks, 1984; Percy, 1986; Rosentraub & Thompson, 1981), whereas other studies have shown no correlation (Brown & Coulter, 1983; Lineberry, 1977; Stipak, 1979).

Table 1
Sample of Recommended Sewer Performance Indicators From Several Sources

Indicator	Description
World Bank (2004)	
Sewerage coverage	Percentage of population with direct service connections
Complaints	Percentage of complaints/total sewer connections/year
Wastewater treatment	Percentage of collected sewage receiving primary treatment
Total operating revenue	U.S. dollars/connection/year and U.S. dollars/household/year
Percentage of gross domestic product	Total annual operating revenue/population served/gross domestic product/capita
Residential fixed charge	U.S. dollars/residential connection/year
Charge ratio	Average charge to industrial users/average charge to residential users
Connection charge	U.S. dollars/residential sewer connection
U.S. Environmental Protection Agency (2002)	
Staffing level	Number of employees per 1,000 customers
Personnel cost	Personnel cost per customer
Sewer revenue	Income from sewer system activities
Blockages	Number of blockages per 10 km of pipe length
Overflows	Number of sewer overflows per 1,000 consumers
Operation and maintenance costs	Operation and maintenance costs per meter of pipe length
Rehabilitation	Percentage of rehabilitated sewers per total pipe length
Complaints	Number of written complaints
Pollution	Number of pollution incidents at sewers
Collapses	Number of sewer collapses per 1,000 km
Overflows	Number of sewer overflows (except in extreme weather)
Renovations	Length of sewers renovated
Replacements	Length of sewers replaced
Total rehabilitations	Total length of sewers renovated and replaced
Internal flooding	Total percentage of properties reporting internal sewage flooding
Installation costs	Sewer main installation costs per kilometer
Rehabilitation costs	Sewer main rehabilitation costs per kilometer
American Society of Civil Engineers (1998)	
Condition	Condition and performance grade
Capacity	Need versus capacity grade
Funding	Funding versus need grade
Yepes (1993)	
Length 1	Kilometers of pipe/number of people served
Infiltration	Gallons per day/inch-diameter of pipe/mile
Length 2	Kilometers of pipe/number of connections
Reader, Tremey, Kempe, Friis, and Garcia-Poveda (1983)	
Coverage	Number of people served
Density	Number of people served/kilometer of pipe
Length	Kilometers of pipe
Janssens, Pintelon, Cotton, and Gelders (1996)	
Consumer complaints	Number of complaints/1,000 connections/year
Number of failures	Number of failures
Number of severe failures	Number of failures taking over 1 day to repair
Top 10 critical list	Ten toughest repair jobs in the time period
Reliability	Time period/number of failures
Maintainability	Time spent on repairs/number of repairs
Breaks	Number of breaks/kilometer of pipe/year
Staffing	Number of employees/1,000 connections
Costs	Annual operation and maintenance costs/population served
Working ratio	Annual operation and maintenance costs/annual operating revenue

Parks (1984) and Kelly and Swindell (2002a) concluded that subjective and objective outcome measures each provide valuable but different information about performance and that both types of empirical data are needed to understand the various dimensions of urban service outcomes. All information from users is not subjective; some information is based on observations and experiences unobtainable from any other source (Percy, 1986). Furthermore, so-called objective measures of performance are often unavailable or inaccurately collected.

Kelley and Swindell (2002a, 2002b) and Van Ryzin (2004) recommended measuring urban service outputs and outcomes at the neighborhood level, because service outcomes often vary considerably from one neighborhood to another. Abbott (1996) used performance indicators that represented the viewpoints of both technical professionals and the community, noting that technical professionals "viewed services in terms of their performance against the original design specification, the quality of the workmanship and materials of construction" (p. 166), whereas "the community uses a more qualitative criterion of value . . . i.e., how well the service performs its task" (pp. 169, 170).

The Evaluation Setting

Projects and programs are evaluated to determine whether they are "meeting their stated goals and objectives, and whether they are doing so with some degree of efficiency" (Schalock, 2001, p. 170). The focus of a performance evaluation may be project effectiveness (i.e., the degree to which the project meets objectives) or project efficiency (i.e., the degree to which project benefits exceed costs). Project effectiveness exists independently of a project's perceived economic value or the efficiency with which a project recovers costs (Abbott, 1996). Both aspects of performance are important and, ideally, should be evaluated concurrently to provide a more complete picture of project outcome. However, there are situations in which an evaluator may choose to focus on one aspect of performance.

Abbott's (1996) evaluation of urban service delivery in South Africa provides a relevant example of performance evaluation based on project effectiveness criteria alone. Abbott encouraged the separation of project efficiency measures from project effectiveness measures in evaluating urban services for the poor. In his research, Abbott found that the criteria used by beneficiaries to evaluate services differed qualitatively from the criteria used by engineers and officials. Beneficiaries valued a project in terms of the service it actually performed, whereas engineers and officials valued the degree to which a project recovered costs and conformed to specifications. This divergence in goals supports a two-pronged approach to evaluating performance, in which efficiency concerns are disentangled from considerations of effectiveness.

In this article, I focus on project effectiveness for three reasons. First, there were no established criteria for evaluating the effectiveness of condominial sewers, so even if projects were justified on efficiency grounds, it was difficult to understand how projects actually performed. Understanding the outcomes and impacts of condominial sewer projects is an important prerequisite to quantifying these benefits in an efficiency evaluation. In fact, assessment of the economic efficiency of a sewer project is predicated on assumed benefits that may not be realized for low-performing projects (i.e., projects with low effectiveness). Second, my objective was to develop a framework for evaluating performance that included different perspectives about project effectiveness. Integrating the perspectives of beneficiaries into the evaluation necessitated a focus on project effectiveness because, as Abbott (1996) concluded, beneficiaries are primarily concerned about effectiveness.

Third, cost data needed to evaluate project efficiency were not readily obtainable because the case studies were implemented in the late 1980s and early 1990s, during an era of severe inflation in Brazil. Brazil has experienced some of the highest inflation of any country, with inflation rates as high as 2,489% in 1993 (Mainwaring, 1999, p. 90). From 1985 to 1994, the average inflation rate in Brazil was 1,018% (p. 90). During that period, the price of a condominial sewer project would be recalculated each month because the value of Brazilian currency was constantly changing. One condominial sewer project manager I interviewed reported that "by the end of [a] project, no one knew the true cost" of any project implemented at that time. Further complicating the estimation of true costs was the government subsidization of construction materials and monthly fees for all sewer services (conventional and condominial).

Given this set of circumstances, I concluded that it was intellectually more defensible to separate the analysis of effectiveness from the analysis of efficiency and that I had the resources and time to perform only one evaluation. Consequently, project effectiveness is the focus of my performance evaluation. The terms *performance* and *effectiveness* are used interchangeably in the remainder of this article.

Condominial sewer performance is mediated by many factors, which results in a complex evaluation setting. In the Brazilian context, one cannot assume that there is adequate knowledge among users about sewer usage, consistent sewer maintenance by the sewer agency, full coverage of all homes with sewer systems, or the ability to prevent the influx of squatters into any neighborhood. The Brazilian context for sewer evaluation is unlike conditions in the United States. For example, in his analysis of the distribution of urban services in San Antonio, Texas, Lineberry (1977, p. 130) found no correlation between the percentage of housing units with public sewers and the socioeconomic status of different census tracts. In Brazil, however, access to basic sewer service is well correlated with socioeconomic status (e.g., Tolosa, 1978). Consequently, an outside evaluator cannot determine performance by site visits or output indicators alone. The following evaluation scenarios illustrate why recommended performance indicators can be inadequate for characterizing condominial sewer performance.

Figure 1 shows the sanitary conditions of a low-income Brazilian neighborhood in which condominial sewers have been installed. (Not visible in Figure 1 is the rest of the neighborhood, which has a relatively well-functioning condominial sewer and few squatters.) The neighborhood is located within the right-of-way of a railroad track and adjacent to the city's wastewater stabilization ponds. Cleanout box covers (Arrows 1 and 2 in Figure 1) indicate the location of underground condominial sewer piping. Note the open ditch (Arrow 3 in Figure 1) that runs parallel to the railroad tracks. Depending on weather conditions, the wastewater in this ditch consists of runoff water, gray water, or sewage. Gray water is waste wash water (e.g., from bathing, dish washing, and clothes washing) that has not been mixed with sewage. A potential source of the gray water and sewage in the ditch is households whose condominial sewers did not function properly, which led people to dispose of their gray water and sewage in the ditch. Poor sewer performance that results in the use of ditches for disposal—whether because of damage to the system, a lack of ongoing maintenance, improper design, or poor construction—can be just as bad for public health as no sewer service at all, even if all of the homes are connected to the sewer system. Heller (1999, p. 140) found that the most commonly used indicator of sewer system performance, the percentage of households connected, showed no statistically significant correlation with health protection for urban sewerage systems in Betim, Brazil. In this scenario, some recommended performance indicators, such as the rate of sewer pipe leakage or the rate of sewer overflows, would provide useful information about the quality of the service, but other indicators, such as the percentage of households connected, would provide an incomplete and possibly inaccurate measure of performance.

In a different scenario, the gray water and sewage in the ditch could be from households that did not connect to the condominial sewer, either because connection was not required at the time the project was implemented or because families could not afford to connect. In this sce-

Figure 1
A Low-Income Brazilian Neighborhood With Condominial Sewers

nario, the sewer may function well but would not achieve the full beneficial impacts intended because of the low connection rate. However, many other recommended indicators of sewer performance, such as the total kilometers of pipe installed or the flow capacity of the pipe, would not effectively measure performance in this scenario.

In another scenario, the gray water and sewage in the ditch could be from households that connected to but did not use the condominial sewer for 100% of their wastewater disposal. In this scenario, social norms were not effective in dictating peoples' use of the condominial sewer, resulting in reduced beneficial impacts, no matter how well the sewer functioned or how many homes were connected. No recommended indicator of sewer performance, including the percentage of households connected or the rate of sewer pipe leakage, would effectively measure performance in this scenario, because performance is mediated by the way people use the system, and peoples' use of the system determines performance, to some degree.

In yet another scenario, the gray water and sewage in the ditch could be from squatters who moved into the neighborhood after the condominial sewer project was installed. Some people find it easier to set up informal residences in low-income or vacant areas whose existing inhabitants may have settled there illegally, which is the case for the neighborhood in Figure 1. Landless and homeless people typically do not have access to sewer systems, and consequently, they dispose of their wastewater in open ditches or other public areas. These survival practices reduce the intended beneficial impacts of a condominial sewer system, no matter how well the sewer functions or how many homes are connected. In this scenario, no recommended indicator of sewer performance would accurately reflect performance outcome. Ultimately, the performance measures selected must be applicable to the setting, diverse enough to account for a number of possible scenarios, and numerous enough to avoid omitting potentially important variables.

In addition, residents' involvement in the production of condominial sewers should inform the evaluation approach. Compared with nonparticipatory projects, community participation can expand the objectives of the sewer system beyond technical criteria. These expanded objectives might include building community capacity to work collectively; educating individual residents about the links between sewage and disease; changing local social and cultural norms about the acceptable levels of disease risk; informing residents about how to build, maintain, and use their part of the sewer system; and increasing peoples' demand for sewer infrastructure. Understanding how well the system meets objectives such as those generated by participatory planning can be achieved by including user perspectives in the evaluation.

Methods

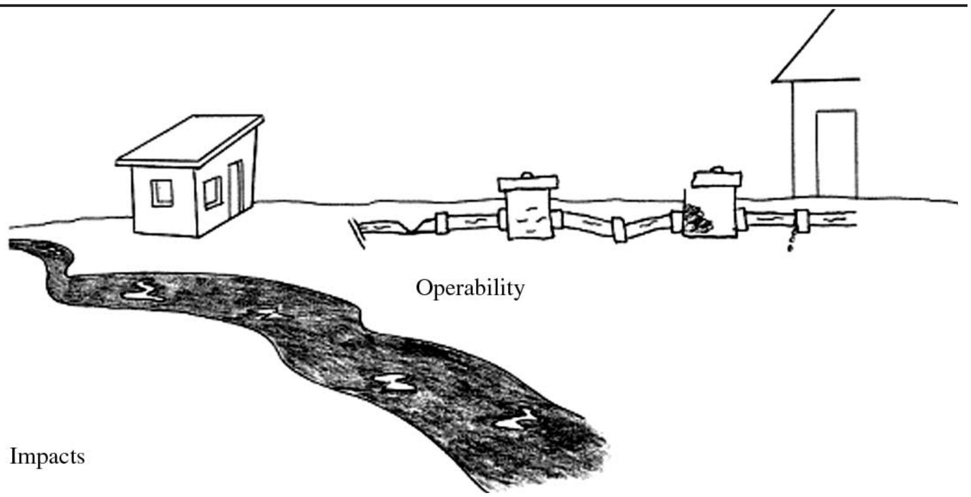
As an outside evaluator conducting research on condominial sewer implementation, my purpose for evaluation was to place individual projects into distinct performance categories so that the cases could be compared with one another and with various independent variables using an inferential process known as "case study with ordinal comparison" (Bennett, 1999). Rather than assess a city's entire sewer infrastructure system, I evaluated a well-defined portion of that system: a distinct project with a unique set of beneficiaries and clear-cut geographical and temporal boundaries, as delineated by local norms. Condominial sewer projects in Brazil are known locally by the name of a major street within the project area, by the name of the housing development in which they are installed, by the name of the neighborhood in which they are constructed, or to a lesser degree by the name of the drainage basin in which they are located followed by the name(s) of the neighborhood(s). The performance of other systems that may have been used to collect, contain, and transport sewage, such as open ditches, storm drains, and street gutters, are not included in the evaluation of the condominial sewer system.

I selected a wide range of condominial sewer performance indicators that drew from the viewpoints of residents, maintenance staff members, engineering staff members, and me as an outside evaluator. The rationale for selecting indicators emerged from my conceptual framework of performance. I constructed the performance variable from two universal objectives: sewer project operability (an output measure) and beneficial impacts (an outcome measure). These universal objectives were adopted because project-specific objectives were not consistently available in documents for the condominial sewer projects studied. In Figure 2, these objectives are presented visually and further described.

I conceived of operability as a measure of how effectively a sewer project performed its basic function, which is to collect, contain, and transport sewage from homes to the discharge point. Fulfilling this basic function requires that each component of a condominial sewer—the house connection, the block sewer, and the street sewer—operate properly. The house connection collects wastewater from household fixtures and removes it from the home, the block sewer collects wastewater from lots and removes it from the block, and the street sewer collects wastewater from blocks and removes it from the project area.

The second objective was beneficial impacts. Impacts are the consequences of a project to residents, the community, and the local environment relative to their expectations, compared with other projects, or relative to a prior condition. The impacts objective included factors such as the number of sewer connections realized, perceived local health and environmental im-

Figure 2
Conceptual Framework of Condominial Sewer Performance



NOTE: Operability represents the functioning and condition of the physical works of a project. If operability is high, then one would expect beneficiaries and sewer agency staff members to give high ratings to project operability, one would expect few reported problems with the system, and one would expect to observe little system damage during a site visit. From these expectations, a list of indicators was derived for measuring the operability of a condominial sewer. Likewise, if a project has high beneficial impacts, one would expect most beneficiaries to report high levels of satisfaction with the sewer, and one would expect residents in the project area to ascribe advantages to the project (as opposed to disadvantages) and be able to identify improvements to local health and environment that resulted from the project. One would also expect to observe relatively clean and dry streets with no public presence of sewage or gray water and a high rate of connection to the sewer among residents in the project area. These expectations formed the basis for establishing a list of indicators for measuring the local impacts of the project.

provements, and how well the project met the needs and aspirations of users. Projects that exhibited high levels of operability and multiple beneficial impacts were characterized as high performers.

The two universal objectives, operability and impacts, established a framework for organizing the numerous performance measures. Project performance, which is the degree to which the two universal objectives were met (otherwise known as project effectiveness), was calculated as the sum of the two objectives, with an implicit assumption of quasi-independence (i.e., that the two objectives were not completely independent or dependent). Although an additive index is presented here, a multiplicative index could equally be applied.

There is some evidence for the quasi-independence of these two objectives. I previously described (Nance, 2004, pp. 143-145) a condominial sewer project (the Grand Favela Project, a pseudonym) that had beneficial impacts on residents, the community, and the local environment even though it functioned poorly. In this case, only the second objective, beneficial impacts, was primarily achieved. In other words, the participation component led to beneficial impacts beyond physical system performance. Some fraction of each objective is usually achieved, but there are also instances in which either objective can dominate. Given results such as this, the two project performance objectives were considered quasi-independent, that is, neither fully dependent nor fully independent.

Table 2
Criteria and Methodology Used to Select Cases

1. Develop a census of all existing condominial projects for each city and a list of agency staff members who know the most about each project.
2. Interview agency staff members one by one, asking each person to identify the best and worst projects in their cities (or to identify the best and worst projects from a list of projects).
3. Visit the named project sites to develop preliminary knowledge about the projects.
4. Compile and analyze the information obtained. Eliminate projects judged best and worst and projects with ambiguous outcomes.
5. Select a set of polar projects that represents both good and bad performance, to the degree possible.

Selecting Case Study Projects

Natal and Recife were selected as research sites because they are each capital cities in Brazil's northeastern region (the poorest area of Brazil). Each city also had more than 5 years of experience with condominial sewers and at least five condominial sewer projects in various stages of implementation. Finally, each city had a variety of project-related information available, as well as officials and agency staff members willing to offer their perspectives and insights. Seven condominial sewer projects in Natal and Recife were selected as case studies.[1] The strategy for case selection was to identify a small number of polar cases that were the "best" and the "worst" projects in terms of performance, on the basis of the judgments of local practitioners (see Table 2). Polar cases are those most likely to perform either well or badly, an attribute which lends construct validity to the performance measure. The study included three projects in Recife (Cases R1, R2, and R3) and four projects in Natal (Cases N1, N2, N3, and N4). Each case was assigned a pseudonym to maintain household and neighborhood anonymity and to comply with the ethical and privacy requirements for research involving human subjects.

Selecting cases is one of the most challenging aspects of a case study research design, because the lack of knowledge about the values of the independent and dependent variables in the cases can lead to a type of systematic error known as selection bias. Because of this, case study analysts often deliberately select cases on specific variables of interest (George & Bennett, 2005, p. 23). The only sources of information available about the population of condominial sewer projects I studied were the perspectives of local practitioners. In the early stages of case selection, I asked local condominial sewer practitioners (primarily engineering staff members) in Natal and Recife to identify the condominial sewers that in their views exhibited the best and worst functional performance in their respective cities. My reliance on practitioners' views provided two safeguards against potential selection bias. First, I needed local experts to inform my initial selection of cases that represented the full range of the dependent variable of interest,[2] because I wanted to be able to categorize cases on the basis of the value of the dependent variable. Second, I needed to provisionally identify cases of known outcomes for the purpose of examining the causal paths that led to a given level of performance within any individual case. This selection strategy would not have been possible without some preliminary knowledge of the cases provided by local practitioners, even though these initial judgments were later superseded by the results of rigorous within-case analysis.

Out of 49 projects, practitioners in Recife judged 10 projects as best and 12 projects as worst. Out of 19 projects, practitioners in Natal judged 10 projects as best and 8 projects as worst. In arriving at informed judgments of the best and worst projects in their respective cities, practitioners drew from their own knowledge, expectations, and experiences with the local range of condominial sewer performance. Practitioners' judgments in both cities were mixed and sometimes conflicting. In Recife, 3 of the projects were judged best and worst by different practitioners. In Natal, 3 of the projects also ended up in both the best and worst performance categories. It was not a straightforward exercise, even for "experts," to identify the best- and worst-performing projects. These difficulties might have resulted from different interpretations of performance as well as differences in practitioners' knowledge about the on-the-ground performance of individual projects. This variability among local practitioners suggests that the concept of performance is ambiguous and requires more deliberate definition and measuring, which underscores the broader purpose of this article.

Even for those projects that exhibited unanimous agreement among practitioners, individual judgments reflected different rationales. Some practitioners believed that performance was "good" if project implementation was completed (i.e., the obstacles to implementation were so great that just getting a project in the ground was considered a success). Other practitioners considered performance as good if a project served a large number of people (i.e., the project achieved a high coverage rate). Other practitioners judged performance as good on the basis of how the sewer system actually functioned (i.e., its physical operability was satisfactory) or on the basis of the beneficial impacts of the sewer system (i.e., improvements to local health and environment were attributed to the condominial sewer system). There was no clear reason why some projects were judged best or worst by different individuals. These results highlight the possibility of high sampling and measurement error if practitioners' judgments are the sole source of project performance information.

The population of projects just described includes projects for which practitioners unanimously agreed on performance (i.e., unambiguous performance) and projects for which practitioners did not agree on performance (i.e., ambiguous performance). A sample of seven unambiguous projects was selected for in-depth study. These seven projects included Cases R1, R3, N1, N2, and N4, judged unambiguously by local condominial sewer practitioners to be among the best-performing projects in their respective cities, and Cases R2 and N3, judged unambiguously by local practitioners to be among the worst-performing projects. In the next phase of data gathering, detailed information about the performance of these selected cases was obtained from multiple stakeholder groups.

Gathering Performance Data

Because of the lack of reliable and available empirical data on condominial sewer performance, I collected a wide range of primary data from four sources: engineering staff members, maintenance staff members, project area residents (including project beneficiaries and non-beneficiaries who lived within the project areas), and direct observation inside homes and around the project areas. (Local officials were interviewed as well, but not for the purpose of acquiring project-specific information.) I limited my use of secondary data to sewer system drawings, from which I obtained project boundaries, the total count of blocks in the project areas, and the total count of street manholes for each project. These secondary data were verified in the field when possible, especially for the smaller projects. Data from residents were designed to be representative of the project areas, so households were selected randomly and representative sample sizes and response rates were sought. Target sample sizes were designed to achieve confidence intervals of ±15%. Actual sample sizes exceeded the targets for all seven cases and resulted in actual confidence intervals ranging from ±10.5% to ±13.5%. Practitioners, on the other hand, were uniquely identified (using snowball sampling methods) rather than randomly selected, because representativeness was not an objective for this group. In total, qualitative and quantitative data were gathered from 297 project area households and 31 practitioners (i.e., engineering staff members and maintenance staff members) across the seven case studies.

Table 3
Indicators of Sewer Performance From Multiple Stakeholders

Indicator	Description	Stakeholder Group
Operability indicators		
HouseOper	1. Ratings of house connection operability	Residents
StreetOper	2. Ratings of street sewer operability	
ConnectProb	3. Reported problems with sewer and house connections	
StreetBlock	4. Reported blockages in the street sewer	Maintenance staff members
CondoBlock	5. Reported blockages in the condominial sewer	
HouseBlock	6. Reported blockages in the house connections	
SoilBlock	7. Reported blockages caused by soil	
TrashBlock	8. Reported blockages caused by trash	
DamageBlock	9. Reported blockages caused by damaged pipes/covers	
SewageBlock	10. Reported blockages caused by sewage	
StormBlock	11. Reported blockages caused by storm water	
PumpProb	12. Reported malfunctioning of the sewage pump station	
PlantProb	13. Reported malfunctioning of the sewage treatment plant	
CoverProb	14. Reported lost or damaged manhole covers	
OperTime1	15. Ratings of condominial sewer operability at Time 1[a]	Engineers
OperTime2	16. Ratings of condominial sewer operability at Time 2	
SealedCovers	17. Observed number of sealed street manhole covers	Direct observations
IntactCovers	18. Observed number of intact household manhole covers	
IntactPipes	19. Observed number of intact household sewer pipes	
BuriedPipes	20. Observed number of buried household sewer pipes	
Impact indicators		
SatisTime1	21. Ratings of satisfaction with condominial sewer at Time 1	Residents
SatisTime2	22. Ratings of satisfaction with condominial sewer at Time 2	
Disadvantages	23. Reported disadvantages of the condominial sewer	
HealthEnviron	24. Reported improvements to health and environment	
HouseConnect	25. Observed number of households with sewers	Direct observations
Graywater	26. Observed frequency of public gray water (dry weather)	
Sewage	27. Observed frequency of public sewage (dry weather)	
Sewer performance index	Median of Operability + Median of Impacts	All of the above

a. Time 1 refers to the time period just after project completion (1988 to 1991). Time 2 refers to the time of the interview (1994 to 1995).

Because each case started with a different set of initial conditions, observations from residents, engineering staff members, and maintenance staff members were essential. Only these individuals were familiar enough with each case study locale to judge differences in the before and after conditions of the neighborhoods. Direct observation was also essential as a consistency check from one locale to another, especially across cities. Also, because each community may have had different aspirations, it was necessary to obtain input from residents on their levels of satisfaction and their perceptions of environmental and health impacts resulting from the projects.

Table 3 presents the 27 direct observations and interview questions that were used to estimate project performance. The operability component of performance was measured by 20 questions and observations, including rating questions (on an ordinal scale ranging from 1 to 5), categorical questions (yes or no), and direct observations that were a mix of categorical, ordinal, and numerical information. The impacts component was measured by 7 questions and observations, including categorical and ordinal questions, rating questions, and direct observations composed of ordinal and numerical data. Using nonparametric statistics, these performance measures can be used to rank order, categorize, and compare projects.

Constructing a Sewer Performance Index

The 27 performance measures were converted into scores, expressed as the percentage of maximum possible on a scale ranging from 0 to 100, for which a score of 100 was most favorable and a score of 0 was least favorable. Scores for interview response data were normalized to the total number of respondents (which was different for each question) and normalized to the range of possible scores per item (to account for categorical, three-scale ordinal, and five-scale ordinal questions). The equation for determining scores from interview responses was

$$\text{Score} = [(R_{\text{int}} - s_{\min}) / (s_{\max} - s_{\min})] \times 100, \text{ where } R_{\text{int}} = [\Sigma(r \times n) / \Sigma n].$$

R_{int} is the average response to the interview question, Σn is the total number of respondents per question, n is the number of responses at value r, r is the value of an individual response, and s is the question scale (typically 1 to 5, 1 to 3, or 1 to 2).

For the numerical (as opposed to categorical or ordinal) data gathered by direct observation, the number of observed instances of a phenomenon was divided by the total number of observations made and multiplied by 100 to convert the data to scores on a scale ranging from 0 to 100. The equation for determining scores from direct observations was

$$\text{Score} = R_{\text{obs}} \times 100, \text{ where } R_{\text{obs}} = \Sigma o / T.$$

R_{obs} is the average observed value, o is the value of an observation, Σo is the sum of the observed values, and T is the total number of observations.

Median scores for each performance indicator (i.e., operability and impacts) were used in a sum-of-weighted-factors equation to calculate a sewer performance index (SPI) for each project. Sum-of-weighted-factors methodology was further described by Loomis (2002, p. 116). The equation for determining the SPI was

$$\text{SPI} = (O \times w_{\text{o}}) + (I \times w_{\text{i}}).$$

O is the median operability score, I is the median impacts score, w is the weighting factor, and $w_{\text{o}} + w_{\text{i}} = 1$. Equal weights (i.e., $w_{\text{o}} = w_{\text{i}}$) were used because I had no basis for weighing one indicator more heavily than another; weights should be assigned by stakeholders. As a result of using equal weights, the sensitivity of the index to any indicator was equal to the degree of variability of that indicator. Although different weights are not used in this article, the inclusion of weights in the index allowed for the possibility of trade-offs between the two objectives. For example, if an evaluator determines that the impacts objective is more important than the operability objective, then this preference (or trade-off) can be operationalized by increasing w_{i}, the impacts weighting factor. The SPI reflects the overall percentage of performance achieved within a range of possible scores for the questions and observations that make up the index.

Results

Overall performance results are summarized in Table 4, in which higher scores represent better performance. Rather than describing the results of each individual performance measure for each case, I describe below a few pertinent examples.

Input from maintenance staff members revealed that the best-performing cases had a low incidence of downstream wastewater treatment plant problems (median score = 71), and the worst-performing case had a high incidence of such problems (score = 0). Plant problems affected the performance of condominial sewers by restricting flow and causing wastewater to

Table 4
Aggregated Performance Scores for the Seven Case Study Projects

	Median Score (maximum = 100)						
Performance Objective	Case R1	Case R2	Case R3	Case N1	Case N2	Case N3	Case N4
Operability	58	7	55	56	48	47	58
Impacts	77	17	87	80	84	63	86
Sewer performance index	67	12	71	68	66	55	72

back up in the system, or by preventing the proper discharge of the collected sewage. Direct observations revealed that all of the best-performing cases had very high percentages of well-sealed manhole covers (median score = 93), and the worst-performing case had a midrange percentage of well-sealed manhole covers (score = 44). Inadequately sealed manholes allowed debris and storm water to enter the system and cause blockages and overflows.

All of the best-performing cases had high levels of beneficiary satisfaction several years after system implementation, as reported by residents (median score = 71). Beneficiaries of the worst-performing case reported low satisfaction several years after implementation (score = 26). Direct observation also revealed a high house-to-sewer connection rate (median score = 78) for all of the best-performing cases and a midrange connection rate for the worst-performing case (score = 48). Also, the best-performing condominial sewer cases had a low incidence of open air sewage in the project areas (median score = 100), and the worst-performing case had a high incidence of open air sewage (score = 0) as determined by direct observation.

Analysis

A nonparametric statistical method known as the median chi-square test of independence was used to arrange the cases into statistically significant categories. A nonparametric test had to be used because the data were a mix of ratio, ordinal, and categorical data that were nonnormally distributed. The median test was selected because unlike other nonparametric tests (e.g., the Mann-Whitney *t* test, the Kruskal-Wallis *H* test, and the Wilcoxon rank-sum test), the median test is appropriate when a nonparametric set of data contains more than a few ties (Hinton, 1995; Huck, 2000), which was the case here. (See Nance, 2004, for more details.) The results of this test showed no statistically significant difference (at $p \leq .05$) among the median performance scores for six of the cases (Table 5). Therefore, these seven cases, when compared against one another, should be placed into two statistically significant performance groups. Cases R1, R3, N1, N2, N3, and N4 make up the higher performance group, and Case R2 makes up the lower performance group.

Can Expert Practitioners Identify Good and Bad Projects?

As previously discussed in "Methods," local condominial sewer practitioners initially judged Cases N3 and R2 as the worst performers and the remaining cases as best. The performance data obtained by in-depth case analysis confirmed that Case R2 was a poorly performing project; however, these data placed Case N3 in the good performance group because Case N3's performance score (SPI = 55), although midrange, was not statistically different from performance scores for the better performing cases (at $p \leq .05$). For Case N3, practitioners' judgments were not congruent with the SPI measure of performance.

Table 5
Results of a Series of Median Chi-Square Tests of Significance on the 27 Unweighted Performance Scores for Each Case

	Median Chi-Square Test[a]			
Test	*df*	χ^2	$\chi^2_{critical}$	Reject Null Hypothesis?[b]
Main test				
Case R1				
Case R2				
Case R3				
Case N1	6	17.1838	12.5916 ($p \leq .05$)[c]	Yes
Case N2				
Case N3				
Case N4				
Post Hoc Test 1[c]				
Case R1				
Case R3				
Case N1				
Case N2	5	2.0282	13.8383 ($p \leq .01667$)	No
Case N3				
Case N4				
Post Hoc Test 2[c]				
Case R1	1	1.5104	5.7308 ($p \leq .01667$)	No
Case N3				
Post Hoc Test 3[c]				
Case R2	1	5.7778	5.7308 ($p \leq .01667$)	Yes
Case N3				

a. The median chi-square test methodology was described by Huck (2000, p. 656).
b. The null hypothesis used for this test was that the sample population medians would all be equal.
c. To maintain an overall level of significance of $p \leq .05$, a Bonferroni adjustment was applied to the three post hoc tests, resulting in a level of significance of $p \leq .01667$ per test (Huck, 2000).

The details of the case provide the information needed to explain this discrepancy. Condominial sewer performance in the Case N3 project area was not homogeneous. Within the Case N3 project area, there was a localized performance problem that affected only a small fraction of project area households (perhaps 10%). Local practitioners might have considered this problem significant, as reflected in their unanimous judgments that this project was among the worst-performing condominial sewers in Natal. It is possible that practitioners judged overall project performance on the basis of the existence of a localized problem. Although the depressed performance score for Case N3 (SPI = 55) does not accurately reflect the significance of the localized performance problem (because the SPI averaged the entire project area), knowledge of the problem emerged from the multistakeholder evaluation process, not from asking practitioners about the best and worst projects.

There are other possible explanations for the discrepancy with Case N3. One possibility is that practitioners considered other information that was not included in the 27 performance indicators, such as the number of complaints or a wider range of weather conditions than was observed during the data collection period. Another possibility is that practitioners in Natal and Recife had different expectations for what constituted bad performance. Practitioners in Natal may have had relatively higher expectations of performance than practitioners in Recife, and this would explain why the practitioners in Natal judged Case N3, a midrange-performing project, as one of the worst in their city. These possible explanations for the discrepant views of Case N3 are supported by data I gathered on this case, on other cases, and on the municipal settings in which the cases were implemented (see Nance, 2004, chap. 6, appendix F). Information about residential complaints was available for Case N3, and the practitioners I interviewed were well aware of these data; however, I did not use these complaint data as a performance indicator because comparable data were not available in Recife. These data could have informed the judgments of local practitioners in Natal and contributed to the discrepancy. Also, I found that Natal and Recife had considerably different levels of overall performance, which supports the possibility that different expectations of performance contributed to the discrepancy.

Practitioners' knowledge, expectations, and experience should have been reflected in the performance ratings made by engineering staff members, but most of the engineers interviewed were unable to rate individual projects on a scale. Response rates dropped sharply as I moved from open-ended interview questions (i.e., in which engineers were asked to identify the best- and worst-performing projects) to rating-type interview questions (i.e., in which engineers were asked to rate the performance of specific projects on a scale ranging from 1 to 5). I interviewed 13 engineers in Recife, but no more than 3 could rate the performance of any particular project. Likewise, I interviewed 12 engineers in Natal, but no more than 4 could rate the performance of any particular project. These results represent an average response rate of only 25% for the engineering staff members I interviewed. By comparison, I interviewed 2 to 4 maintenance staff members per project in Recife and 4 maintenance staff members per project in Natal, and I obtained an average response rate of 84%. These results suggest that engineers in the two cities studied were not readily able to rate the performance of individual condominial sewer projects, even though they had definite opinions about the best and worst condominial sewer projects in their respective cities.

The results show that initial performance judgments by engineering practitioners were not entirely reliable. The value of practitioner judgments was in identifying, within a single city, what might be the best and worst performers for case selection purposes, but this information had to be verified by in-depth case analysis, especially for projects that lacked homogeneous performance.

Do Members of a Stakeholder Group Agree on Performance?

To estimate the degree of agreement within each group of stakeholders, the 27 performance measures were disaggregated by group (residents, engineering staff members, and maintenance staff members). For the purposes of estimating the degree of agreement among respondents, direct observations were not included as a stakeholder group because the research design did not include multiple independent observers. The level of agreement within each group was estimated in two ways. For the engineering and maintenance staff member groups, in which a relatively small number of practitioners were interviewed for each case, I counted how frequently a simple majority of respondents (i.e., >50%) had the same response to the performance questions associated with that group, expressed as a percentage (i.e., the number of simple majorities divided by the number of questions). For the residents group, in which representative numbers of randomly selected households were interviewed, I performed a chi-square goodness-of-fit test (at $\rho \leq .05$) on the following null hypothesis: The sample population exhibits no preference for any single category. I counted how frequently this null hypothesis was rejected for the resident performance questions, expressed as a percentage (i.e., the number of rejections divided by the number of questions).

Of the three stakeholder groups, maintenance staff members had the highest degree of agreement in their responses to 11 questions about project performance. For Cases R2, R3, N1, N2, N3, and N4, the frequency at which a simple majority of maintenance staff members had matching responses ranged from 91% to 100%. For Case R1, a simple majority had matching responses for 64% of the questions.

Table 6
Median Performance Scores (Disaggregated by Stakeholder Group) and Mean Average Deviation (MAD) Medians for Each Case

Stakeholder Group	Case R1	Case R2	Case R3	Case N1	Case N2	Case N3	Case N4
Operability							
Engineers	65	0	57	53	41	41	67
Maintenance staff members	50	0	25	33	25	0	50
Residents	48	13	54	59	54	53	44
Direct observations	98	79	96	100	97	100	100
Impacts							
Residents	54	33	74	83	73	77	72
Direct observations	100	0	100	77	96	50	100
MAD median							
All groups	19	21	23	20	24	25	26

Note: The MAD median is the mean average deviation from the median performance score. A MAD median of 0 would indicate that all stakeholder groups assigned the same score. A MAD median of 50 would indicate maximum variability in the scores assigned by stakeholder groups.

The resident group had the next highest degree of agreement in their responses to seven questions about project performance. For Cases R1, R3, N1, N2, N3, and N4, the frequency at which the null hypothesis was rejected ranged from 71% to 86%. For Case R2, the null hypothesis was rejected for 57% of the questions. Overall, a majority of residents responded similarly to most of the performance-related questions for all seven cases.

The engineer group had the lowest degree of agreement in their responses to two questions about project performance. For Cases R1, N2, N3, and N4, a simple majority of engineers had matching responses to one of the two questions (frequency = 50%). For Case R2, a simple majority had the same response to both questions (frequency = 100%). For Cases R3 and N1, a simple majority was not achieved on either of the questions (frequency = 0%). The relatively low degree of agreement among engineering staff members was striking, especially if one considers that engineers are typically the primary group of practitioners involved in project evaluation research. Methodologies that rely exclusively on “expert” ratings of performance by engineers may have to be reconsidered in light of these findings.

Do Perspectives Differ Across Stakeholder Groups?

To assess whether the perspectives of different stakeholder groups varied significantly compared with one another, median performance scores were disaggregated by group (see Table 6). Medians were used rather than averages because of nonnormality in the distribution of scores, nonuniform sample sizes and response rates, small sample sizes, and the lack of interval-level data for all of the data points. The degree of variability across stakeholder groups was estimated as the mean average deviation (MAD) from the median performance score. MAD medians for all stakeholder groups ranged from 19 to 26 (out of a maximum variability of 50 and a minimum variability of 0) for the seven cases studied. These results indicate a relatively high degree of variability across stakeholder groups; that is, different stakeholders scored the same project differently. This was expected and further validates the rationale for including multiple stakeholders. But these results do not indicate any pattern in stakeholder variability between cases with poor and good performance, because there was little difference in the MAD medians across the seven cases.

Two patterns do emerge from the data in Table 6. First, for all seven cases, operability scores from maintenance staff members were among the lowest compared with operability scores from residents, engineering staff members, and direct observations. This suggests that maintenance staff members were more critical of condominial sewer performance or that perhaps maintenance staff members had higher expectations than the other stakeholders. Second, for all seven cases, operability scores from direct observations were consistently higher and less variable than operability scores from residents, engineering staff members, and maintenance staff members. This suggests there were limits to what could be learned from observing the operability of the selected condominial sewer systems. For the cases studied, either the kinds of physical conditions that could be observed (i.e., the conditions of manhole covers and sewer pipes) did not vary widely between cases with poor and good performance, or poor performance was due to factors that did not result in visible damage (such as inadequate design or improper use).

Discussion

Multistakeholder evaluation was superior to evaluation based on practitioner judgments and was both complementary and superior to evaluations based on recommended indicators such as those listed in Table 1. Because the multistakeholder approach elicited and triangulated objective and subjective performance data from several relevant stakeholders, it provided a more robust view of performance than a single stakeholder group could provide. In addition, the multistakeholder approach separated the cases into distinct performance categories more reliably than practitioner judgments, even though the practitioners selected (most often engineers) were well experienced and were in a position to categorize projects in terms of performance.

Although the engineers I interviewed were purposively selected, it is possible that they did not have up-to-date knowledge about the ongoing performance of projects they had planned, designed, or implemented. For example, some condominial sewer projects in Natal and Recife were judged best and worst by different practitioners. Furthermore, Case N3 illustrated how the evaluation of performance was confounded by localized failure within the boundaries of an otherwise well-performing project. On the basis of practitioners' judgments, Case N3 was originally categorized as a worst-performing project, but the multistakeholder approach revealed that Case N3's overall performance was comparable with that of the best-performing projects in both cities. Multistakeholder evaluation was, therefore, superior to the sole use of practitioner judgments.

For the purpose of placing projects into performance categories, the use of direct observations was complementary to evaluation methods based on recommended performance indicators. As would be expected in many developing-country settings, performance data were either not available or not taken at the project level in the two cities studied. In lieu of these data, I gathered independent data by making direct observations of system operability and system impacts in the selected case study project areas. The data obtained from these direct observations, which included quantitative information about the conditions of sewer pipes and manhole covers, household connection rates, and the incidence of open air sewage and gray water within the project areas, were proxies for some of the data typically used to derive recommended indicators.

There were limits to the use of direct observation results, however. Unlike instrument measurements that can provide data year round and cover a citywide sewer system, direct observations took place in a relatively short time period, over a limited range of weather conditions, and in neighborhood-level portions of the sewer system. Also, direct observations of system operability were found to be somewhat less discriminating than direct observations of system impacts in terms of separating projects into performance categories. However, the practice of directly observing project level performance was an effective approach for obtaining independent quantitative information and for identifying local failures within a neighborhood or project area.

The multistakeholder evaluation approach used in this research reveals that performance is an ambiguous concept. Even among projects largely considered successful, perspectives on performance varied from individual to individual within a single stakeholder group. This was predominantly true for the engineer group, which had the highest level of within-group variability among the stakeholder groups studied. Performance is also a contested concept, because when the perspectives of different stakeholder groups were included, the resulting information about performance was not always congruent. In most of the cases studied, even those with the best performance, maintenance staff members' performance scores were considerably lower than scores of engineering staff members and residents. Consequently, the pursuit of only one group's perspective on performance would likely skew the results. A superior approach is multistakeholder evaluation, which triangulates sources to improve the validity and defensibility of performance evaluation. By combining independent direct observations with information from different stakeholders, multistakeholder evaluation provides a more complete picture of performance than recommended indicators alone.

Summary of Findings

Expert judgments of the performance of individual projects were useful in the initial identification of polar cases but were found to be unreliable for cross-city comparisons and for projects with heterogeneous performance. Direct observations of system impacts differentiated project performance more decisively than direct observations of system operability. Performance evaluations by maintenance staff members were consistently more critical than those of other stakeholders, and as a group, maintenance staff members had high internal agreement. Project area residents also displayed high internal agreement in their evaluations of performance. Compared with the other stakeholders studied, engineering staff members had the lowest internal agreement about performance and the lowest response rates.

The median chi-square test of independence was a practical nonparametric method for statistically confirming the assignment of cases to performance categories. The use of the SPI was useful for summarizing project performance and comparing cases from different contexts. However, analyzing disaggregated performance measures was also important, because it provided detailed information about the level of achievement on each of the 27 performance measures, it revealed the degree of agreement within each stakeholder group, and it highlighted differences among stakeholder group views of performance.

Lessons Learned

Many improvements and refinements to the multistakeholder evaluation procedure are possible. If I were to conduct this research again, there are several ways I would improve the methods for gathering performance data. I would use a common scale for more of the interview questions to facilitate coding and analysis. To address the lack of household-level public health data, I would conduct interviews with doctors from local health clinics for information regarding sanitation-related disease. I would also redouble my efforts to identify a larger sample of engineers who are knowledgeable about individual projects, and I would intensify the search for neighborhood-level complaint data.

Conclusions

This article presents an evaluation of condominial sewer projects in a developing-country context using a multistakeholder approach. Prior to this article, condominial sewer technology had been evaluated at the city level in the international development literature, input from multiple stakeholders had been used to evaluate the performance of urban services in the public administration literature, and recommended indicators for sewer system evaluation had been presented in the engineering literature. Unique to this article is a conceptual framework for organizing the performance indicators, the evaluation of neighborhood-level projects using information from multiple stakeholders, the development of the SPI, and the use of the median chi-square test to arrange projects into statistically significant categories. The results in this article show that recommended performance indicators cannot be relied on to measure condominial sewer performance because of the complexity of the evaluation setting (e.g., the developing-country context, the expanded project objectives generated by participatory planning, and the lack of project-level performance data). The article further shows that expert judgments of individual project performance were not always reliable, and that different stakeholders had different views of performance. The adoption of a multistakeholder approach will help practitioners and evaluators develop evaluation programs that are meaningful to all relevant stakeholders and will contribute to growing agreement about how condominial sewers should perform and how their performance should be measured.

These findings have a broader significance to the field of evaluation because they inform our understanding of concept validity. Concept validity depends on "the utility of the device that measures it and the collective judgment of the scientific community that a construct and its measure are valid" (Bernard, 1994, p. 43). This article demonstrates why evaluators cannot always assume that recommended indicators represent an underlying concept in a manner with which all relevant stakeholders agree. Evaluators cannot always assume that recommended indicators represent a collective viewpoint; the indicators might instead represent a contested viewpoint or the viewpoint of a single stakeholder group. Moreover, evaluators cannot always assume that indicators proved to be valid in one context will retain their utility in a different context.

The fundamental lesson to be drawn from this article is that in complex evaluation settings with wide-ranging views toward evaluands, input from multiple stakeholders is needed to identify the relevant dimensions of a concept and ultimately to develop valid operational constructs. This lesson leads to two questions that evaluators should ask: (a) Are existing operational constructs valid and relevant in the current evaluation context? and (b) Do existing operational constructs encompass diverse views of the evaluand? If the answer to either of these questions is negative, then a multistakeholder evaluation approach might be warranted.

Notes

1. Detailed case study descriptions are presented elsewhere (Nance, 2004).
2. I also relied on practitioners' judgments to inform case selection on the key independent variable (i.e., community participation), as described elsewhere (Nance, 2004).

References

Abbott, J. (1996). *Sharing the city: Community participation in urban management*. London: Earthscan.

American Society of Civil Engineers. (1994). *Existing sewer evaluation and rehabilitation* (2nd ed.). Reston, VA: Author.

American Society of Civil Engineers. (1998). *Report card for America's infrastructure*. Retrieved November 23, 2004, from http://www.asce.org/reportcard

Bennett, A. (1999). *Causal inference in case studies: From Mill's methods to causal mechanisms*. Retrieved November 19, 2004, from http://www.georgetown.edu/faculty/bennetta/APSA99.html

Bernard, H. R. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Birckmayer, J. D., & Weiss, C. H. (2000). Theory-based evaluation in practice: What do we learn? *Evaluation Review*, *24*(4), 407-431.

Brown, K., & Coulter, P. B. (1983). Subjective and objective measures of police service delivery. *Public Administration Review*, *43*(1), 50-58.

Brudney, J. L., & England, R. E. (1982). Urban policy making and subjective service evaluations: Are they compatible? *Public Administration Review*, *42*(2), 127-135.

Chelimsky, E., & Shadish, W. (Eds.). (1997). *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: Sage.

Fetterman, D. (1997). Empowerment evaluation: A response to Patton and Scriven. *Evaluation Practice*, *18*(3), 253-266.

George, A., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Heller, L. (1999). Who really benefits from environmental sanitation services in the cities? An intra-urban analysis in Betim, Brazil. *Environment and Urbanization*, *11*(1), 133-144.

Hinton, P. R. (1995). *Statistics explained: A guide for social science students*. London: Routledge.

Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.

Janssens, J. G., Pintelon, L., Cotton, A., & Gelders, L. (1996). Development of a framework for the assessment of operation and maintenance (O&M) performance of urban water supply and sanitation. *Water Supply*, *14*(1), 21-33.

Kelly, J. M., & Swindell, D. (2002a). A multiple-indicator approach to municipal service evaluation: Correlating performance measurement and citizen satisfaction across jurisdictions. *Public Administration Review*, *62*(5), 610-621.

Kelly, J. M., & Swindell, D. (2002b). Service quality variation across urban space: First steps toward a model of citizen satisfaction. *Journal of Urban Affairs*, *24*(3), 271-288.

Lineberry, R. L. (1977). *Equality and urban policy: The distribution of municipal services*. Beverly Hills, CA: Sage.

Loomis, J. B. (2002). *Integrated public lands management: Principles and applications to national forests, parks, wildlife refuges, and BLM lands*. New York: Columbia University Press.

Mainwaring, S. (1999). *Rethinking party systems in the third wave of democratization: The case of Brazil*. Stanford, CA: Stanford University Press.

Mohr, L. B. (1999). The impact profile approach to policy merit: The case of research grants and the university. *Evaluation Review*, *23*(2), 212-249.

Nance, E. B. (2004). Putting participation in context: An evaluation of urban sanitation in Brazil. *Dissertation Abstracts International*, *65*(04), 2035B. (UMI No. 3128354)

National Council on Public Works Improvement. (1988). *Report card on the nation's public works*. Retrieved November 23, 2004, from http://www.asce.org/reportcard

Ostrom, E. (1973). The need for multiple indicators in measuring the output of public agencies. *Policy Studies Journal*, *2*(1), 85-91.

Ostrom, E. (1996). Crossing the great divide: Co-production, synergy, and development. *World Development*, *24*(6), 1073-1088.

Parks, R. B. (1984). Linking objective and subjective measures of performance. *Public Administration Review*, *44*(2), 118-127.

Percy, S. L. (1986). In defense of citizen evaluations as performance measures. *Urban Affairs Quarterly*, *22*(1), 66-83.

Reader, N. F., Tremey, G., Kempe, A., Friis, P., & Garcia-Poveda, M. (1983). Performance indicators. *Water Supply*, *1*(2/3), SS21-1-SS21-4.

Rosentraub, M. S., & Thompson, L. (1981). The use of surveys of satisfaction for evaluations. *Policy Studies Journal*, *9*(7), 990-999.

Schalock, R. L. (2001). *Outcome-based evaluation* (2nd ed.). New York: Kluwer Academic.

Stipak, B. (1979). Citizen satisfaction with urban services: Potential misuse as a performance indicator. *Public Administration Review*, *39*(1), 46-52.

Tolosa, H. C. (1978). Causes of urban poverty in Brazil. *World Development*, *6*(9/10), 1087-1101.

U.K. Office of Water. (1998). *Level of service indicators*. Retrieved November 23, 2004, from http://www.ofwat.gov.uk/aptrix/ofwat/publish.nsf/Content/losindicators

U.S. Environmental Protection Agency. (1977). *Sewer system evaluation, rehabilitation, and new construction: A manual of practice* (No. EPA-600/2-77-017d). Cincinnati, OH: Author.

U.S. Environmental Protection Agency. (1991). *Sewer system infrastructure analysis and rehabilitation* (No. EPA/625/6-91/030). Cincinnati, OH: Author.

U.S. Environmental Protection Agency. (2002). *Decision-support tools for predicting the performance of water distribution and wastewater collection systems*. Cincinnati, OH: Author.

Van Ryzin, G. G. (2004). Expectations, performance, and citizen satisfaction with urban services. *Journal of Policy Analysis and Management*, *23*(3), 433-448.

Watson, G. (1995). *Good sewers cheap? Agency-customer interactions in low-cost urban sanitation in Brazil*. Washington, DC: World Bank.

World Bank. (1994). *World development report 1994: Infrastructure for development*. Washington, DC: Oxford University Press.

World Bank. (2004). *Benchmarking water and sanitation utilities: Network of core indicator values*. Retrieved October 18, 2004, from http://www.worldbank.org/html/fpd/water/topics/bench/bench_network_inddef_ac.html

Yepes, G. (1993). *Water and sanitation utilities: Operational indicators*. Washington, DC: World Bank.