Chapter 5

A New Index of Telephone Service Quality: Academic and Regulatory Review

Sanford V. Berg
University of Florida

1. INTRODUCTION

The academic ideal, for many scholars, is to do research that is simultaneously theoretically elegant and solves important practical problems in the real world. In this chapter, I provide a case history of my own attempt to pursue this ideal. My work focused on the regulatory measurement and reward of service quality provided by regulated local telephone service companies operating in the state of Florida. I describe the challenges my research team faced in (a) maneuvering a relatively straightforward regulatory innovation into regulatory practice, and (b) disseminating the conceptual framework to another community (through publication in scholarly journals read and edited primarily by academics).

Although others will evaluate the ultimate success of this dual research strategy, there are lessons to be learned from the effort. Generalizations from one observation are problematic—but even one data point can shed light on the issues associated with the adoption of rules and procedures which improve resource allocation in telecommunications. This work suggests that both sound theory and supporting empirical evidence are necessary if current approaches to quality are to be strengthened. I offer some low-brow theory and observations on actual regulatory behavior to draw lessons regarding the role of academic research in promoting regulatory innovations.

2. OVERVIEW OF CURRENT REGULATORY SCHEMES AND THE PROPOSED INDEX

The importance of service quality has been highlighted by developments in the last decade: divestiture, network interconnection, and technological change (Rovizzi & Thompson, 1992). The competitive and complementary service offerings of new entrants raise challenges for incumbent local exchange carriers. In the meantime, state regulators are still faced with the choice between traditional cost-of-service regulation and various forms of incentive regulation. Whatever their decision, the role of quality has a higher profile than in the past.

The regulatory process utilizes technical performance features of networks, even while recognizing that consumer satisfaction may depend only indirectly on engineering measures of service quality. These surrogate evaluations tend to consist of pass/fail technical standards that were often established decades ago. They have grown by accretion: The most recent National Association of Regulatory Utility Commissioners (NARUC) compendium on the subject lists between 90 and 100 separate standards (depending on how one groups some subcategories).

Current quality-of-service pass/fail targets are somewhat arbitrary, having arisen from a chaotic process reflecting historical engineering capabilities, political pressures, and administrative happenstance. Consumer valuations of different quality dimensions and corporate recognition of emerging technological opportunities are not likely to be captured by pass/fail standards. In addition, combining information on multiple dimensions into an overall assessment is very difficult for regulators.[1] Information overload could lead to "management by exception." By focusing on the rules that a company fails, regulators essentially ignore dimensions on which the company being evaluated has exceeded the standards. Similarly, the use of cutoff targets gives companies currently operating below a standard no incentive to improve performance if those improvements would still leave them short of the standard. Thus, perverse incentives result from the use of pass/fail standards. Developing an appropriately weighted quality-of-service index is no simple task, but the approach represents a potential improvement over multiple pass/fail quality standards.

2.1 Production Possibilities for Pass/Fail Standards

Current reward schemes, in Florida as in other states, compare a company's objective scores, $Z_1 \ldots Z_n$, on a set of engineering attributes to standards, $Z^*_1 \ldots Z^*_n$, set by the regulatory agency on those attribute dimensions. Below-standard performance on any attribute triggers censure, whereas performance above any standard is not treated as being any better than performance exactly at the standard. This tacitly drives companies to produce quality along each attribute at exactly Z^*_i, because exceeding Z^*_i generally takes resources but is not rewarded

in the regulatory system. Beyond this, however, very little could be said about exactly what regulators were rewarding or trying to reward, as they used subjective and intuitive judgment to assess the overall service quality of a firm that had a complex set of above-standard and below-standard scores on the many measured dimensions of quality.

We attempted to make their expert judgment more systematic by modeling regulators' trade-offs among various dimensions of quality, $\hat{Q} = f[(Z_1 - Z_1^*), \ldots, (Z_n - Z_n^*)]$. We discovered that regulators often agreed that overall quality was higher when Standard A was exceeded and B was failed (when Attribute A was considered relatively more important) than when A and B were met exactly. Therefore, existing evaluation policies were creating perverse incentives by treating the former company as being in violation and the latter as being in compliance with regulations.

In our proposed alternative index, we define $\hat{Q}^*$ to be the predicted level of overall quality associated with meeting all n quality standards exactly. We then propose that any combination of $Z_1 \ldots Z_n$ that leads to $\hat{Q} \geq \hat{Q}^*$ should be treated as meeting the standard. Under this regime, companies would offer whatever combination of substandard and superstandard $Z_1 \ldots Z_n$ that allows them to achieve $\hat{Q}^*$ in the most efficient way given that company's cost structure. For each 1% point above a standard, the quality index rises—depending on the weight given to that particular standard. Similarly, shortfalls result in reductions in the index.

For simplicity, consider two dimensions of service quality monitored by regulators: dial-tone response (Z_1) and call completions (Z_2). Like other commissions, the Florida Public Service Commission (FPSC) has standards for each of these. Florida requires that 95% of all calls receive a dial tone within 3 sec. The intraoffice call completion standard requires the successful completion of 95% of all calls to numbers with the same first three digits as the calling number. A welfare-maximizing regulator will induce the firm to equate the marginal benefits from service quality improvements with the marginal costs. This condition for optimality is depicted in Fig. 5.1 (adapted from Berg & Lynch, 1992). Three production possibility frontiers (PPFs) are shown—those that are further out require that additional resources be devoted to the production of quality: \$100, \$110, and \$130, respectively. For a given level of real resources, improvements in one quality dimension involve a deterioration in the other. The PPF reflects engineering and resource constraints. The slope of the PPF represents the opportunity cost of increasing Z_2: Given the constraints, there must be a reduction in Z_1.

2.2 Relative Valuations of Service Characteristics

Relative valuations for the two dimensions of service quality are also shown in Fig. 5.1. In this example, the subjective trade-offs by customers are reflected in the preference mapping characterized by $U = 2{,}010$ and $U = 2{,}020$. That is, for any given level of satisfaction (for example, $U = 2{,}010$), if one dimension de-

teriorates (say Z_2—call completions—falls from 95% to 90%), then Z_1 must increase if customers are not to be made worse off. Here, Z_1 (dial-tone response) must rise from 95% to 97% if the customers are to remain on $U = 2{,}010$.

In this example, points E and M represent the same level of satisfaction, met by different combinations of service qualities. Point E would not be a welfare maximizing point because point M is valued equally by consumers and costs less to achieve. At point E, the subjective marginal rate of substitution between Z_1 and Z_2 does not equal the marginal rate of transformation (as reflected in the slope of the production possibility frontier). The proposed approach would drop the pass/fail standards of 95%, 95%, and give the telephone company flexibility in selecting least-cost ways to achieve a given level of performance. In the example, point X could be achieved for the same resource cost as point E—but benefits would now be $U = 2{,}020$.

Figure 5.1 illustrates how firms could be presented with a regulatory objective function and allowed to trade off high-cost (low-valued) quality dimensions for low-cost (highly valued) quality dimensions. In the simple example, if (95,95) yielded an "acceptable" overall level of quality, one scoring function that would signal the telco to modify its quality mix would be $Q = Z_2 + (5/2)Z_1$, and the minimum quality "score" is $Q = 332.5$. The firm would go to point M (97, 90).

One issue is whether the minimum quality "score" is appropriate. In the simple example, if $Q = 337.5$, the firm will be driven to point X instead of point M. If point M corresponds to \$105, and an additional \$5 lets the firm attain point X, then the outlay is worth it if X is valued \$5 or more than the quality bundle at M. We did not try to attack the incremental cost/incremental valuation issue, but focused on replacing the (95,95) standard with a minimum quality score of 332.5. The question is how to derive the weights described earlier.[2]

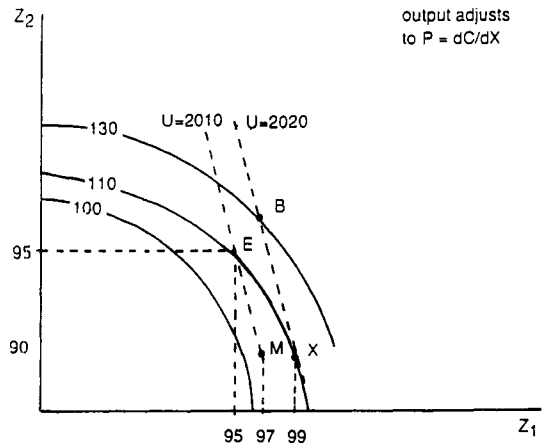

FIG. 5.1. Relative valuations of quality attributes.

This is an extremely simple idea that can be easily shown to represent an improvement over the current regulatory regime, so long as there is agreement as to the weights to be given the various standards (Noam, 1991). Yet, the legal, political, and institutional roadblocks to its full adoption and appropriate implementation have been numerous. At the same time, we have published and presented this work in several academic venues. However, the primary theoretical piece is still battling to emerge from a 4-year review process at a prestigious academic journal that strives to merge theory and practice.

3. SCIENTIFIC REVIEW AND THE CREATION OF KNOWLEDGE

Publication lags, like regulatory lags, arise from the existence of numerous checkpoints in which stringent review criteria are applied. Academic gatekeepers provide critical reviews of analyses, giving readers some confidence that the published article represents a contribution to the literature.[3] External reviewers can help researchers focus their efforts and remedy potential flaws in analyses. The review process screens potential contributions, asking whether the submission contributes new and creative insights regarding the issue at hand. A second and perhaps more fundamental question is whether the issue under consideration is actually important.

Theoretical constructs, empirical tests, and historical evaluation provide the three legs upon which a policy science stands. Given the gains from specialization and division of labor, researchers will tend to tackle issues from one of the three perspectives, although good analysis using any one of the three modes cannot ignore the other two. For example, good theory recognizes the historical setting that establishes the institutional context for theoretical analysis. In addition, theory often depends on empirical observations (in the form of stylized facts) to provide bounds on key parameters or to determine the signs of particular relationships. Reviewers of potential contributions know the economic paradigm from which models are derived and quantitative tests conducted. Deviations from the widely accepted neoclassical economic framework face a hurdle in the review process. Because this work does not draw heavily on the paradigm of the rational, evaluative, maximizing consumer, I have had to justify the framework to skeptical academic reviewers.

The concern here is with the evaluation of telecommunications service quality. The literature on product (or service) quality is voluminous. Because the theory is summarized elsewhere (Berg & Lynch, 1992), it will not be surveyed here. Suffice it to note that the models are elegant, often insightful, and difficult to test. Quality outcomes under competition, monopoly, and regulation depend on a host of factors, including incremental costs, incremental benefits, and average benefits associated with quality changes. Furthermore, the introduction of multiple dimensions of quality greatly complicates the analysis.

Nevertheless, the neoclassical paradigm suggests that informed consumers will evaluate alternative service offerings and select consumption bundles based on their preferences. Thus, the ultimate judge that matters is the rational customer, not some regulatory surrogate. I agree. However, policy analysis cannot abstract from the institutional context: The past matters. Regulatory reviews are designed to screen for a different set of problems than those that might concern scientists. My work has tried to be responsive to both review processes. Indeed, the studies have benefited from discussions with academics and regulators. Nevertheless, it is somewhat risky for academics to adopt a research strategy that tries to meet criteria from both review processes.

4. REGULATORY REVIEW AND PROCEDURAL FAIRNESS

The regulatory review process is grounded in procedural fairness. Disruptive transitions are costly to buyers and sellers alike, so administrative delays can sometimes serve as mechanisms for smoothing out the impacts of unanticipated changes in demand or technologies. Review lags ensure that proposed changes are understood by all those affected by new regulations. In addition, administrative procedures are designed to provide opportunities for complaints to be heard. Hearings and informal workshops serve as forums in which stakeholders present their concerns.

The heavy role of legal, accounting, and engineering expertise at public utility commissions suggests that economics by itself provides an inadequate foundation for regulatory decisions. Emphasis on legal precedent gives continuity to the rate-making process—forcing some consistency in the face of emerging problems. Protection is afforded both the regulated firm and its customers. From the standpoint of service quality, fairness toward customers requires that data be verifiable. If reported data are fabrications or the result of improper manipulation of data collection procedures, the firm loses credibility. Credibility is essential if the regulatory process is to be accepted by consumers.

Similarly, the heavy dependence on accounting data constrains much of the debate to the consideration of actual outlays, rather than hypotheticals. Although future test years are utilized in many jurisdictions, the focus is still on accounting rather than economic costs. Economic opportunity costs may be considered, but accounting data and cost allocation procedures based on historical developments dominate rate cases so long as there are no alternative suppliers. In addition, engineering data are particularly relevant for considering quality-of-service issues because these are objective and subject to review.

The regulatory concern is that regulated (or partially regulated) firms may choose the wrong level and mix of quality, recognizing that quality is multidimensional. But what does *wrong* mean? Too little quality? Too much quality?

An inappropriate mix of quality components? An inappropriate pricing of quality components?

The framework described here is appropriate for encouraging the right mix of quality characteristics. By itself, it does not address the overall level of quality. However, the current battery of pass/fail standards addresses neither the mix nor the level of quality. By building on current data collection efforts, the proposed approach maintains continuity. In addition, the framework allows regulators to formalize what they mean by quality. The procedure for soliciting weights has its limitations. However, the weights can be refined. With the adoption of a quality index, managers can make network investment and operations trade-offs based on precise weights applied to a specified set of characteristics. This represents an improvement in the process because the current pass/fail targets are subject to uneven regulatory application. The current process can be quite inefficient as well as potentially unfair.

Thus, from the standpoint of procedural fairness, the proposed quality index is quite promising. As with any new instrument, however, telecommunications firms will be hesitant to accept new rules. For example, two problems with the movement to price caps are the determination of the starting price level and the calculation of the productivity adjustment. Both would have to be determined in advance before a telco would support the replacement of rate of return on rate-base regulation with price caps. Similarly, before firms will accept a new quality index, they will want to know how it is to be used in the regulatory process. If it is seen as another tool for bludgeoning the firm, the index will not receive their support. Strong opposition by regulated firms reduces the likelihood of adoption.

Intervenors will also be skeptical of any departures from current conventions. For example, the Public Counsel's Office will primarily be representing residential customers. The weights appropriate for these customers may not be the same as those for high-volume commercial and industrial demanders. Because the dimensions of quality tend to be collective consumption goods (quality available to one is available to all), consumer advocates may not want to give the firm the discretion involved in meeting an overall quality-index constraint. Rather, they may prefer to focus on items of particular concern to their constituency.

Given these observations regarding the focus on continuity and aversion to change, stakeholders will delay adoption of new instruments by regulatory agencies until they have made thorough checks on implications for performance. Regulators, utility managers, and consumer advocates will all need to be comfortable with the new approach.

5. CASE STUDY OF A NEW QUALITY INDEX

The rationale behind the proposed quality index, Q, has already been described. Lessons can be learned by reviewing the parallel evolution of the conceptual framework and associated regulatory rules. Let us begin at the beginning: The

investigation was initiated in late 1986, when FPSC staff approached the Public Utility Research Center (PURC) about exploring ways to evaluate quality of service provided by local exchange companies. Based at the University of Florida, PURC attempts to bridge principles and practice, so the issue clearly fits into its purview.

5.1 Initial Data Collection

Academic institutions are not consulting groups who can switch resources rapidly from one activity to another. The author tried to identify researchers who could assist with the project. Two marketing professors had ideas about how a more comprehensive indicator of quality might be developed. Interactions with FPSC staff yielded what was thought to be a good understanding of the pass/fail standards applied to telcos. Teaching responsibilities and other research commitments meant that much of the initial investigation occurred in the summer of 1987. Figure 5.2 provides a time line for the stages of regulatory and academic reviews associated with the proposed framework.

The basic methodology involved a survey in which experts made comparisons of quality bundles. The weighting scheme was developed by having experts from the FPSC rate different hypothetical company profiles of performance on the rules within nine rule clusters. Each profile was rated on a scale from 1 (worst possible performance) to 10 (best possible performance). Similar comparisons were made across rules so that a comprehensive score could be assigned to a telco based on its observed performance on the 38 dimensions. The entire procedure is a form of conjoint analysis called the hierarchical conjoint analysis (Louviere, 1984). (See the Appendix for an example of how weights were derived.) This approach is suited to capturing the trade-offs experts make in overall evaluations of objects that can differ on a very large number of attributes that can be logically grouped into subsets of related attributes.

In January 1988, a report was ready for the FPSC (Buzas & Lynch, 1988). The methodology for determining weights to be given the various dimensions was outlined in some detail, and associated statistical tests were presented. It provided illustrative calculations for hierarchical conjoint analysis so the derivation of individual weights could be described. The study analyzed how agreement and disagreement among survey participants could be identified. The report included a discussion of each item on the questionnaire, defined technical terminology, and showed which of the 38 dimensions had greatest weights, based on responses provided by FPSC staff. The formula derived showed the weight of a 1 percentage point change on each of the 38 dimensions.

The three-member research team prepared an academic working paper in July 1988. This paper was the first project output focusing on both the methodological and policy issues associated with the new index. The weights were calculated and applied to a hypothetical telephone company. The implications of agreement

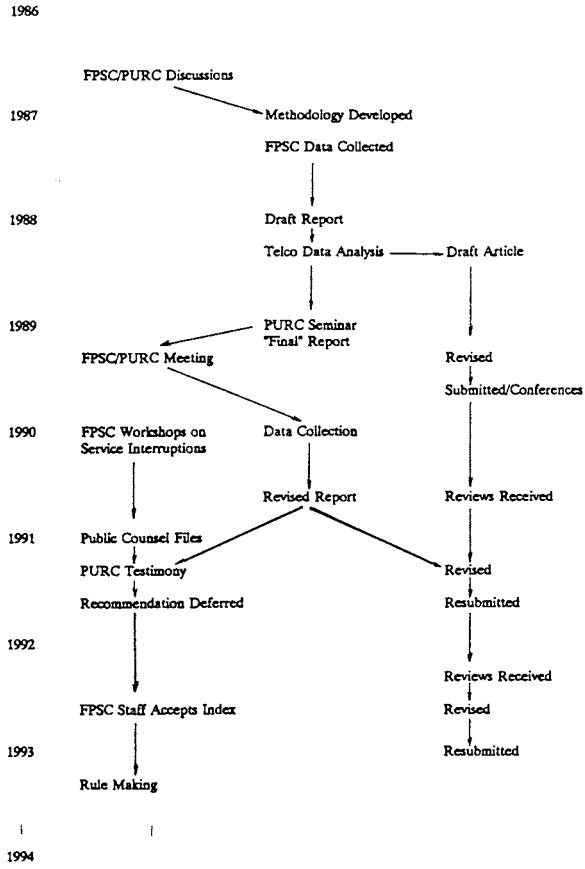

FIG. 5.2. Regulatory and academic reviews.

and disagreement among survey participants were also outlined. The report contained examples of techniques for determining the reliability of estimates.

Finally, limitations to the study were identified. In particular, the separate role of costs was discussed. Also, because the initial trade-offs were made by FPSC employees, it was noted that these experts might have a different perspective than either consumers or companies. Third, the trade-offs were made without reference to specific locations or clientele. It is plausible that the FPSC might want to reward compliance on some rules more heavily or lightly in certain geographic areas. For example, compliance on public telephone dimensions might be more important in rural areas than in urban areas because the phones are further apart in the former.

The research team needed to determine whether experts at telephone companies would give the same weights to the various rules. We thought that telco representatives might be more aware of the relative benefits of meeting the

different rules. Alternatively, different firms face different mixes of customers (due to demographics, per capita income, degree of urbanization). Such factors could mean that customer valuations for different components of quality differed across firms, making a single weighting scheme inappropriate. Despite the remaining issues, the team believed that the results were highly suggestive: Progress was being made in this very complicated area. The research team viewed the methodology as offering a way to introduce greater rigor and content into the quality evaluation process. The project turned to the issue of telco weights.

The data collection effort moved forward in earnest, as the team sought cooperation from regulated firms. Some were willing to devote personnel to the effort, but others were concerned about implementation issues. A PURC seminar was held in February 1989. Berg, Buzas, and Lynch described the rationale and methodology behind the comprehensive index. Results from the FPSC sample (12 employees) and two companies were presented. Alan Taylor, Chief of the Bureau of Service Evaluation, represented the FPSC. Also in attendance were representatives from the major Florida telcos. These representatives expressed a concern as to how the proposed index might be used: to evaluate firms at a single point in time? to evaluate trends over time for a single firm? to compare firms? The different service territories and degree of network modernization influence the starting point for each firm, raising a concern for fairness. Cross-firm comparisons might not take into account different technological opportunities. Executives tended to see the index as another factor that could be used against them at the next rate case.

Most of the formal presentations focused on basic methodological points:

1. Different dimensions had different weights.
2. Only a few dimensions really mattered a great deal.
3. There was a very high correlation (Pearson correlation coefficient of better than .90) between weights obtained from the two companies and from the commission.

A hypothetical example was given, and the method of calculation was presented. To illustrate the usefulness of a comprehensive index, data for two hypothetical companies were presented, emphasizing the difficulty of making pass/fail comparisons. The politics of regulation was not given much attention, although attendees were probably more worried about implementation issues than statistical refinements.

The initial scoring function had 38 weights plus a base "score" if each standard as exactly achieved:

$$Q_a = 5.92 + .1172(Z_1 - Z_1^*) + .0786\,(Z_2 - Z_2^*) + .0813(Z_3 - Z_3^*) + \ldots + .0198\,(Z_{38} - Z_{38}^*)$$

Here $Z_1^* = 95\%$ of calls that received a dial tone in 3 sec

$Z_2^* = 95\%$ of intraoffice (same first three digits) calls completed

$Z_3^* = 95\%$ of interoffice calls completed

•

•

•

$Z_{38}^* = 100\%$ of all public telephones that have their locations posted, and the identifications of locations coordinated with the appropriate 911 or emergency center.

For this example, a telephone company just meeting each standard would score 5.92. A company score of 96% on dial tone delay (Z_1) would more than offset a company score of 94% in interoffice call completions (Z_2), in which the standard is 95% for each but the weight for the former is greater than the weight for the latter. Note that 16 of the 38 weights in this initial scoring function referred to public telephones: functioning (receives calls), enclosures (handicapped access, cleanliness, and lights), coin operations (coin returns, operator assistance), directory availability, and so on. In addition, the linear form of the scoring function implied that a 1 percentage point improvement for a service dimension had the same impact whether the change was from 90% to 91% or from 96% to 97%. This was approximately true regardless of whether the change represented a movement toward the standard or one which exceeded the standard.

Within weeks, the Final Report on Telephone Service Quality was sent to the FPSC. The academics thought their jobs were done. Some project results were reported in a conference volume that appeared 2 years later (Buzas, Lynch, & Berg, 1991). After further work, the team submitted "Regulatory Management and Evaluation of Telephone Service Quality" to *Management Science* in late 1989. This analytical study attempted to bridge principles and regulatory practice. In addition, a review paper on service quality was presented at the Telecommunications Policy Research Conference and the Southern Economic Association meetings that fall. This second working paper was directed to a mix of academics and technically trained policymakers. It reviewed the literature and described the work on the quality index consolidating the 38 characteristics. After being rejected by a theoretical journal, a revised version received favorable reviews at *Telecommunications Policy*, in which it was published in early 1992 (Berg & Lynch, 1992).

5.2 Adapting to Reviews

While still waiting for initial academic reviews of the two manuscripts, the team obtained formal reactions from the FPSC on the "final" report. In a letter, Gene Ferguson, a FPSC engineer, identified a number of deficiencies in the proposed evaluation weighting system:

> I would suggest that when the PSC experts are again chosen, they select those of us who understand the rules, procedures, essentials of traffic switching, maintenance and maintenance terms, network structure and network switching, business office and repair facility operation, and the effects of any deficiencies on the subscriber, either directly or indirectly and to the rate payer.

It seems clear the Bureau of Service Evaluation felt that it was not adequately represented in the initial survey! Furthermore, Ferguson was concerned with the wording of the questions and the number of quality dimensions omitted from the initial survey. Lack of initial input from FPSC technical engineers clearly put off the staff. Yet team members thought they were going through correct channels and had no inkling that key stakeholders (in this case, commission engineering staff) had not been utilized, either in developing the questions or taking the survey.

For example, the term *NR* reported in FPSC evaluation reports are often interpreted by companies as *No Rule*. However, the *NR* means that *N*o specific percentage *R*equirement was spelled out in the rule. Often 100% compliance is required (when the word *all* is implied). In other cases, the FPSC selects realistic targets for these rules. The initial study did not include these standards.

Also, when staff applied the weighting scheme to a recent evaluation in which a company had failed to meet pass/fail standards in 19 areas, the index score was above average. Although the particular calculation involved a misapplication of the methodology, the example raised red flags regarding the implications of switching to a single index. In addition, in the case of "same-day restoral" (omitted in the initial survey), the FPSC requires 80%. Because this was perceived as an important service standard, they wanted this item included. A number of other omitted items were identified. To illustrate omissions, the team discovered that FPSC engineers had three standards related to central office exchange facilities. In addition, there were seven transmission rules. For example, an exchange with a dial-tone level outside the range of −5 to −22 dBM is not in compliance with the standard. Ferguson wanted some weight given to these items.

A subsequent meeting with FPSC staff in Tallahassee cleared up a number of misunderstandings (on both sides). Follow-up telephone conversations between Buzas and Ferguson helped each understand the others' concerns. The result was a memo by Buzas and Lynch that addressed data collection, aggregation, and evaluation. They raised a number of points.

1. Is the index a gross or net measure of relative benefits? Because service quality is conceptually distinct from cost of service, the focus here was on the benefit side. The trade-off against cost could be handled separately from the relative valuation of quality dimensions. The concern had been expressed that the adoption of a comprehensive index could lead to goldplating. The key point was that for many dimensions, improvements above the standard yields incremental benefits that are greater than the associated incremental costs.[4] Furthermore, the incentives to provide excessive levels of quality were no different than at present.

Just meeting every standard might be more costly than reaching another set of performance targets that yield an identical comprehensive score. If the weights are correct, just meeting each standard is goldplating in the sense that costs are too high for achieving this overall level of quality. Current FPSC mechanisms for evaluating network modernization programs and other prudency tests would apply to the cost side. At least with a single index, the task of the commission (and firms) would be simplified. The FPSC could focus separately on evaluating the additional costs associated with achieving higher quality scores.

2. *Should a firm pass overall when failing several pass/fail standards?* Between 1985 and 1988, companies failed to reach a standard 191 times. Of these, 162 (or 84.8%) were on the 21 (of 38) rules with a standard of 100%. On average, a company failed 13.6 rules of which 11.6 were on rules with standards of 100%. This suggests that the complexity of evaluating performance by 38 (or more) pass/fail standards is problematic when the degree of "substandard" performance is not captured in the summary index. The relative importance of the failures warrants attention.

The fact that a company can fail numerous rules yet receive a "passing" grade is an integral part of the proposed scheme. The research team argued that setting a numerical passing score does not necessitate a degradation in overall quality. Higher overall scores *could* be required (although such ratcheting up ought to be justified in terms of low incremental costs or high incremental benefits). For example, the lowest passing score achieved by any company might be established as the lower bound, or the average passing score could be taken to be the standard. A minimum acceptable performance rule would be established for each rule, whereas overall performance could be gauged on the basis of a higher level than if all such rules were met exactly.

3. *Do the weights reflect a narrow constituency?* We became sensitized to the likelihood that the FPSC had multiple constituencies to protect. Tourists and nonsubscribers rely heavily on public telephones. Residential subscribers are less dependent on public telephones and so would weight less heavily those dimensions associated with public telephone performance. Similarly, businesses that are using digital transmissions would place a premium on other aspects of quality.

It would be possible to develop separate indices for each constituency. Multiple scores could be reported and evaluated, or these could be weighted to obtain an index of aggregate performance. The team did not view such complications as presenting problems. At least the proposed methodology forces decision makers at commissions and companies to acknowledge that different dimensions of quality performance have different weights.

4. *How does the exclusion of important rules from the weighting formula affect its usefulness?* As it turned out, omitted rules tended to be those without set standards for performance. The initial FPSC liaison was asked about this issue prior to developing the initial survey. The liaison was later sent the questionnaire for prior approval. Unfortunately, that person was in the process of leaving the

commission. Lesson: When trying to interact with regulatory agencies, be sure you have the "right" contact point. Academics might not fully appreciate the need to have an onsite advocate for the methodology (or conceptual framework) under consideration. The exclusion of important rules hurt the initial index's credibility.

5. *Did the right experts complete the survey?* Because the initial liaison did not distribute the survey instrument to key technical personnel with knowledge of the standards, the commission weights were brought into question. However, the weights obtained from the telephone company panel of experts were statistically the same as those obtained from the FPSC. It was clear, however, that a new survey had to be developed and given to additional FPSC staff as well as to cooperating telcos.

6. *How were the formulae to be calculated?* As with any new methodology, a clear understanding of its components was essential for its successful application. In the initial FPSC application, incorrect weights were given to several rules. Also, there was a misunderstanding regarding calculation of the overall index. These points could be easily addressed by preparing clearer instructions, including background information and illustrative calculations.

In summary, the exchange of memos and meetings increased the research team's awareness of the administrative processes used to evaluate quality. Furthermore, the official FPSC response to the conceptional framework remained positive. J. Alan Taylor, Chief of the Bureau of Service Evaluation, noted in a letter that he found:

> the analysis to be an excellent and insightful approach to the problems facing regulators who base their quality of service decisions on simple pass/fail rule criteria. Certainly the regulated industry has long been apprehensive of a service quality measurement regime which focuses primarily on failures, without giving some consideration to more economical improvements in overall performance levels. PURC's weighting system therefore appears to be an appropriate way to assure that general quality of service levels remain high, particularly as we move into an era of regulatory flexibility through an incentive approach to governmental oversight.

The team was anxious to address concerns, while directing attention to the advantages to have a comprehensive quality index. The FPSC decided to fund a new survey and the development of new weights. Lynch and Buzas began the new study.

5.3 New Data and Further Regulatory Adaptations

As the new data collection effort proceeded in early 1990, reviews on the comprehensive academic paper were received. Major revisions were required. The most important criticism was that the team should have asked end-users (con-

sumers) what their trade-offs were rather than asking expert regulators and managers to give their views of the trade-offs that would be in consumers' interests. The team had described theoretical reasons to expect that customers of regulated monopolies would not have established trade-offs—because they have no occasion to choose—but reviewers were unconvinced. The team argued that only if they have expertise are consumers likely to have established trade-offs.

Also, the framework had been presented to the academic and national regulatory community in Fall 1989. Eli Noam served as a discussant of the paper at the Telecommunications Policy Research Conference. As a Commissioner for the New York Public Service Commission and as a fellow academic, he supported the framework, but questioned the linearity assumption. He saw no need to have a symmetry of overperformance to underperformance.[5] However, if one more person's call does not go through, it is not clear why the customer's loss depends on whether or not the standard is already exceeded. Nevertheless, it was agreed that more careful empirical investigation of unchanging weights was warranted.

By late Fall 1990, Lynch and Buzas had a draft of the new report, including a LOTUS 1-2-3 spreadsheet that allowed the FPSC to easily calculate a company's overall quality score using the weights uncovered by the project. The Executive Summary was for practitioners. Technical supporting material was included so that an independent expert could revise the weighted index.

The rule clusters in the second study focused on residential service. Besides retaining 16 public telephone rules, additional rules were added, yielding a new set. In Study 2, the weight given to 1 percentage point improvement leading up to the standard was found to be significantly higher than the weight for 1 percentage point improvements exceeding the standard. Two weights were now utilized:

$$W_{\downarrow} \text{ if } Z_1 < Z_1^*$$
$$W_{\uparrow} \text{ if } Z_1 > Z_1^*.$$

Thus, the concerns of Florida's PSC staff and New York's Commissioner Noam were supported by the new empirical results. For example, a telephone company had its score for dial-tone delay increased by .111 for each percentage point in excess of the standard, but the score fell by .2133 for each percentage point below the standard. It is possible that technical staff completing the second survey were implicitly factoring in the costs of exceeding standards. In another change from Study 1, technical staff from a major customer of telephone services (Florida's Department of General Services—DGS) completed the new survey instrument, allowing a check on the comparability of FPSC and user weights.

In December 1990, FPSC initiated rule making. The rule-making process turned out to be long and tortuous. Ultimately, Lynch testified on the methodology in mid-1991. However, a glitch appeared that further delayed things. The index got caught in regulatory cross-fire. In February 1991, the Public Counsel's

Office filed complaints that a telco had falsified quality-of-service reports. The company was alleged to have falsified records on out-of-service repairs (a pass/fail quality standard). There are two related standards: 80% for same-day repairs; 95% for within 24 hours. The latter is viewed by some as particularly tight. Corporate officials have denied that they encouraged the falsification of records, although technicians may have felt some pressure to do so.

What actually happened is still under debate (and delaying a rate case). The key point is that the regulatory process can get choked through strategic maneuvering by key stakeholders. The Public Counsel's Office did not like the incentive plan that the FPSC had previously adopted for the telco to begin with. The office was unconvinced that rate payers would benefit from the plan. The falsification report provided a wedge for attacking the plan because quality of service was part of the incentive plan: Rate payers would be due refunds. Staff recommendations were deferred.

Thus, if the FPSC replaced the host of pass/fail standards with a single index, some of the steam would taken out of the Public Counsel's case. The telco might argue "mitigating circumstances" because the old rules had been jettisoned. With tens of millions of dollars at stake, there seemed no reason to replace the many standards with a single quality index. Ironically, had the weighted standard been in effect, the alleged undue pressure to falsify might have been a moot point.

In July 1992, rule making was initiated again; a meeting was scheduled for September 22 on Rule No. 25-4.080 (Weighted Measurement of Quality of Service) with the following FPSC announcement:

> The purpose of this new rule is to introduce another tool which the Commission can utilize in its effort to accurately measure the quality of service provided by local exchange telephone companies. . . . The rule authorizes the Commission to utilize a weighted index system when considering the adequacy of service. . . . The system contains various quality of service measures currently contained in Commission rules and weights them according to their importance in the provision of acceptable quality local telephone service. . . . Companies shall be responsible for complying with each service standard, whether or not an overall score of seventy-five (75) is achieved when the weighted index is employed.

An overall score of 75 corresponded to just meeting each standard. The proposed rule has yet to make it back to the Commissioners for a final vote.[6] If accepted, the new rule will use the weighted index as an additional tool for evaluating quality. In contrast to the research team's recommendations, it will *not* replace the dozens of pass/fail standards.

One other development substantiated initial telco concerns that the new quality index may be used to tighten standards. The Public Counsel's Office opposed the index because some of the standards with high weights were relatively easy to attain. FPSC staff considered raising dial-tone delay of less than 3 sec from 95% to 98.5% of the time. Similarly, call completion standards might

be raised from 95% to 98%. Thus, if the index were adopted, the effective pass/fail standards could also be tightened.[7] Furthermore, some pass/fail standards have not been adopted as formal "rules," but if the Commissioners adopt the index, these standards might become pass/fail rules as well. This development would add to the quality constraints imposed on firms. So the formula and its components may become a bargaining point in the regulatory process. This should have come as no surprise and partly explains why changes in the status quo will be opposed by key stakeholders.

Sample calculations for a telephone company are shown in Table 5.1. In this particular case, the absence of information on directory assistance billing accuracy caused a deduction of over 7 points—but this was more than compensated for by surpassing standards in other categories.

5.4 Academic Revisions

At about the same time as the second rule making was initiated (1992), the reviews on the *Management Science* resubmission arrived. These comments required significant additional research (which was already being undertaken to address issues raised by the FPSC). Different issues arose related to consumer perceptions and valuations. Academic reviewers were still concerned that expert regulators rather than consumers filled out the survey instrument. In anticipation of these concerns, the research team had obtained a sample from the largest purchaser of telephone services within the state government—the Department of General Services (DGS). Thus, these respondents were sophisticated and representative of consumers. The team found close agreement between the weights of experts within the FPSC to those of these other experts.

The initial methodology provided a measure of quality; but if quality improvements lead to higher costs to rate payers, the team was unable to say whether such improvements should be encouraged or discouraged. The team had suggested that this would require regulators to marry the system for measuring the marginal benefits of quality with a detailed study of the marginal costs of improvement along various dimensions. Although the spirit of the weighted index was to get the regulators out of the business of micro-managing regulated companies, it was understood that the cost side warranted attention.[8]

The second addition to the initial study extended the work to address this issue. The revised survey instrument elicited trade-offs between various dimensions of quality and the price of basic monthly service, allowing the team to quantify the dollar value of a 1 percentage point improvement along the various dimensions of service quality. Thus, regulators could encourage a quality improvement that would be coupled with a $1 increase in price if the increase in consumer benefits associated with the improvement exceeds the disutility associated with the price increase. Interestingly enough, the FPSC experts gave a dollar value associated with quality increases that was *three times* that obtained from the large demander (DGS).

TABLE 5.1
WEIGHTED INDEX

ST. JOSEPH TELEPHONE CO. REPORT DATE: APRIL 10, 1992
DATES STUDIED: NOVEMBER 11 THRU DECEMBER 13, 1991

CRITERION	*FPSC STANDARD*	*COMPANY RESULTS*	*WEIGHT FACTORS*	*DIFF*	*WEIGHT ADJUST*
A. DIAL-TONE DELAY					
DIAL-TONE DEL +	95.0	100.0	1.1377	5	5.6884
DIAL-TONE DEL –	95.0		8.4935		
B. CALL COMPLETIONS					
INTRAOFFICE +	95.0	100.0	0.0613	5	0.3063
INTRAOFFICE –	95.0		4.0136		
INTEROFFICE +	95.0	100.0	0.0947	5	0.4735
INTEROFFICE –	95.0		2.1075		
EAS +	95.0	100.0	0.0280	5	0.1402
EAS –	95.0		0.9953		
INTRALATA DDD +	95.0	99.9	0.1286	4.9	0.6300
INTRALATA DDD –	95.0		1.0999		
C. INCORRECTLY DIALED CALLS					
INCORRECTLY DIALED +	95.0	100.0	0.1043	5	0.5214
INCORRECTLY DIALED –	95.0		0.1043		
D. 911 SERVICE					
911 SERVICE –	100.0	100.0	2.8772		
E. TRANSMISSION					
DIAL-TONE LEVEL –	100.0	100.0	0.0002		
CENTRAL OFFICE LOSS –	100.0	100.0	0.0002		
M.W. FREQUENCY –	100.0	100.0	0.0002		
CEN. OFF. NOISE METAL –	100.0	50.0	0.0002	–50	–0.0118
CEN. OFF. NOISE IMPLSE –	100.0	87.8	0.0002	–12.2	–0.0029
SUBSCRIBER LOOPS +	98.0	100.0	0.2788	2	0.5577
SUBSCRIBER LOOPS –	98.0		0.1394		
F. POWER AND GENERATORS					
POWER & GENERATORS –	100.0	100.0	0.0798		
G. TEST NUMBERS					
TEST NUMBERS –	100.0	100.0	0.0010		
H. CENTRAL OFFICE					
SCHEDULED ROUTINE PROG +	95.0	100.0	0.0487	5	0.2433
SCHEDULED ROUTINE PROG –	95.0		0.0487		
FRAME +	95.0	100.0	0.0549	5	0.2743
FRAME –	95.0		0.0549		
FACILITIES +	95.0	100.0	0.0758	5	0.3790
FACILITIES –	95.0		0.0758		
I. ANSWER TIME					
OPERATOR +	90.0	99.1	0.0519	9.1	0.4725
OPERATOR –	90.0		0.3820		
DIRECTORY ASSISTANCE +	90.0	97.0	0.0519	7	0.3635
DIRECTORY ASSISTANCE –	90.0		0.3820		
REPAIR SERVICE +	90.0	100.0	0.0519	10	0.5192
REPAIR SERVICE –	90.0		0.3820		
BUSINESS OFFICE +	80.0	98.2	0.0604	18.2	1.0994
BUSINESS OFFICE –	80.0		0.4191		

(Continued)

TABLE 5.1
(Continued)

T. JOSEPH TELEPHONE CO. REPORT DATE: APRIL 10, 1992
ATES STUDIED: NOVEMBER 11 THRU DECEMBER 13, 1991

RITERION	*FPSC STANDARD*	*COMPANY RESULTS*	*WEIGHT FACTORS*	*DIFF*	*WEIGHT ADJUST*
. ADEQUACY OF DIR. AND DIR. ASSISTANCE					
DIRECTORY SERVICE –	100.0	100.0	0.0887		
NEW NUMBERS –	100.0	100.0	0.0399		
NUMBERS IN DIRECTORY +	99.0	100.0	0.2507	1	0.2507
NUMBERS IN DIRECTORY –	99.0		0.5640		
⁽. ADEQUACY OF INTERCEPT SERVICES					
CHANGED NUMBERS +	90.0	100.0	0.1287	10	1.2865
CHANGED NUMBERS –	90.0		0.3107		
DISCONNECTED SERVICE +	80.0	100.0	0.0489	20	0.9775
DISCONNECTED SERVICE –	80.0		0.2151		
VACATION DISCONNECTS +	80.0	100.0	0.0322	20	0.6434
VACATION DISCONNECTS –	80.0		0.0586		
VACANT NUMBERS +	80.0		0.0277		
VACANT NUMBERS –	80.0		0.2079		
DISCONNECTS NON-PAY –	100.0		0.1650		
.. TOLL TIMING AND BILLING ACCURACY					
INTRALATA BILL ACC. +	97.0	100.0	0.4290	3	1.2869
INTRALATA BILL ACC –	97.0		2.8560		
DIR. ASSIST. BILL ACC. +	97.0		0.4794		
DIR. ASSIST. BILL ACC. –	97.0	0.0	0.0766	–97	–7.4277
M. PUBLIC TELEPHONE SERVICE					
1 PAY PHONE/EXCHANGE –	100.0	100.0	0.0006		
SERVICEABILITY –	100.0	92.0	0.0864	–8	–0.6910
HANDICAPPED ACCESS –	100.0	96.0	0.0112	–4	–0.0449
GLASS +	95.0	100.0	0.0056	5	0.0278
GLASS –	95.0		0.0056		
DOORS +	95.0	100.0	0.0051	5	0.0254
DOORS –	95.0		0.0051		
LEVEL +	95.0	100.0	0.0076	5	0.0379
LEVEL –	95.0		0.0062		
WIRING +	95.0	100.0	0.0060	5	0.0298
WIRING –	95.0		0.0141		
CLEANLINESS +	95.0	100.0	0.0005	5	0.0024
CLEANLINESS –	95.0		0.0362		
LIGHTS –	100.0	92.0	0.0224	–8	–0.1793
TELEPHONE NUMBERS –	100.0	100.0	0.0523		
NAME OR LOGO –	100.0	100.0	0.0008		
DIAL INSTRUCTIONS –	100.0	92.0	0.0864	–8	–0.6910
TRANSMISSION +	95.0	96.0	0.0266	1	0.0266
TRANSMISSION –	95.0		0.0266		

(Continued)

TABLE 5.1
(Continued)

ST. JOSEPH TELEPHONE CO. REPORT DATE: APRIL 10, 1992
DATES STUDIED: NOVEMBER 11 THRU DECEMBER 13, 1991

CRITERION	*FPSC STANDARD*	*COMPANY RESULTS*	*WEIGHT FACTORS*	*DIFF*	*WEIGHT ADJUST*
DIALING +	95.0	100.0	0.0008	5	0.0040
DIALING –	95.0		0.0062		
COIN RETURN AUTO –	100.0	96.0	0.0037	–4	–0.0147
COIN RETURN OPER +	95.0	96.0	0.0178	1	0.0178
COIN RETURN OPER –	95.0		0.0178		
OPERATOR ID COINS +	95.0	96.0	0.0002	1	0.0002
OPERATOR ID COINS –	95.0		0.0302		
ACCESS ALL LD CARRIERS –	100.0	100.0	0.0024		
RING BACK OPERATOR +	95.0		0.0002		
RING BACK OPERATOR –	95.0	92.0	0.0302	–3	–0.0905
COIN FREE ACCESS OPER –	100.0	100.0	0.0097		
COIN FREE ACCESS D.A. –	100.0	100.0	0.0042		
COIN FREE ACCESS 911 –	100.0	100.0	0.0093		
COIN FREE ACCESS R.S. –	100.0	100.0	0.0034		
COIN FREE ACCESS B.O. –	100.0	100.0	0.0027		
DIRECTORY –	100.0	100.0	0.0013		
DIRECTORY SECURITY +	95.0	100.0	0.0510	5	0.2551
DIRECTORY SECURITY –	95.0		0.0510		
ADDRESS/LOCATION –	100.0	92.0	0.1252	–8	–1.0013
N. AVAILABILITY OF SERVICE					
3-DAY PRIMARY SERVICE +	90.0	100.0	0.0333	10	0.3332
3-DAY PRIMARY SERVICE –	90.0		0.2406		
PRIM. SERV. APPOINTMNT +	95.0	100.0	0.1306	5	0.6528
PRIM. SERV. APPOINTMNT –	95.0		0.8125		
M. REPAIR SERVICE					
RESTORED-SAME DAY +	80.0		0.0909		
RESTORED-SAME DAY –	80.0	71.2	0.1319	–8.8	–1.1603
RESTORED-24 HOUR +	95.0	100.0	0.3685	5	1.8427
RESTORED-24 HOUR –	95.0		1.3348		
REPAIR APPOINTMENTS +	95.0	96.8	0.1318	1.8	0.2372
REPAIR APPOINTMENTS –	95.0		0.1936		
REBATES OVER 24 HOURS –	100.0	100.0	0.0523		
SERVICE AFFECTING-72 HRS +	95.0	100.0	0.1318	5	0.6590
SERVICE AFFECTING-72 HRS –	95.0		0.1936		
P. CUSTOMER COMPLAINTS	ST. AVE				
COMPLAINTS/1000 LINES +	0.42	0.36	0.3685	0.1333	0.0491
COMPLAINTS/1000 LINES –	0.42		0.0000		
BASE SCORE IF ALL STANDARDS ARE MET EXACTLY			75.00		75.00
SUM OF ADJUSTMENTS					9.00
OVERALL WEIGHTED SCORE (BASE + SUM OF ADJUSTMENTS)					84.00

The average for the FPSC and DGS implied that a $1 change in price was worth .1401 points.[9] If performance on the dial-tone delay standard improved from 97% to 98%, the associated change in the point score was .111 (on a 10-point scale). Thus, a 1 percentage point improvement is worth .111/.1401/$ = $0.79 per month to the customer. This application of the framework to benefit cost analysis represents a potentially valuable extension of this methodology.

5.5 Customer Perceptions and Expert Trade-Offs

It is useful to underscore the role of experts in the team's framework. The second study and associated revisions of academic papers attempted to address the key conceptual issues raised by reviewers.[10] When expert regulators assess service quality in the interest of everyday consumers, the tacit assumption is that the experts' utility functions are highly correlated with those that everyday customers would have if they did not lack knowledge of how measurable attributes translate into realized benefits. Another justification for using experts is that research indicates that when trade-offs among attributes do not already exist in respondents' heads, the trade-offs they construct on the spot are highly unstable. There is a substantial literature showing that experts make trade-offs that are much less sensitive to distorting effects of measurement. The relevant issue is not whether the consumer has experience with the product. The critical concern is whether consumers have experience making trade-offs among the particular dimensions relevant to the decision at hand. Experience means that these trade-offs can be retrieved rather than constructed at the time of measurement (Feldman & Lynch, 1988; Fischhoff, 1993).

The team's approach is supported by Fischhoff (1993), who surveyed a large body of applied policy research that uses "contingent valuation" methodology. That research attempts to elicit citizens' values in the hope that spending on public policy programs can reflect the priorities of consumers. The repeated finding is that when consumers' trade-offs are elicited for goods that are not customarily traded in any marketplace, consumers do not have articulated values relevant to those decisions. The result is that measured trade-offs have indefensible properties.

One could still argue that consumers should identify the criteria to be measured, and that the role of regulators might be circumscribed to judging the quality of credence attributes—those characteristics that are rarely learned, even after consumption. In fact, this process is very close to what actually happens. Consumers call in their complaints to companies and regulators. Such data are reported. The complaints are invariably phrased in terms of benefits. Legal constraints force regulators to reverse engineer those complaints and to trace them back to problems on attributes that underlie those benefits. The problem is that when laws and rules have been written pertaining to these objective attributes (e.g., percentage of interoffice calls completed), they become credence attributes

for consumers. Typical consumers do not understand the links between technical attributes and benefits.

When comparing expert regulators and novice consumers, it is expected that the former generate weights that would be strongly related to those elicited from a trained representative sample of consumers. Consistency across companies, the FPSC, and a large state agency that buys telecommunications services support the research team's view that the weights provide a good first cut at prioritizing the dimensions of quality.[11]

6. CONCLUDING OBSERVATIONS

What was learned? The regulatory process is run by lawyers, with the aid of accountants and engineers. Economists have input into the process, but the deference is underwhelming. Perhaps, this is appropriate. After all, as a profession, precious little has been developed that helps decision makers identify and reward quality.

The research team was pleased that the multiyear research project resulted in a proposed rule for adopting the weighted index system for the evaluation of service quality. However, the proposed rule utilizes the index as an added requirement rather than as a replacement indicator of pass/fail performance standards. Because the spirit of the recommendation was to move away from detailed consideration of the 60-plus dimensions of quality, it is hoped that the FPSC ultimately adopts the comprehensive performance index. If the FPSC utilizes the crude, unweighted pass/fail mechanism as well as the comprehensive index during the transition period, that is their judgement call as to what is politically acceptable.

Are there other lessons from this case study? The multiyear project represents an attempt to serve two masters: one interested in implementing regulatory policy and another interested in extending the boundaries of science. Table 5.2 lists some of the lessons learned in the process. Regulatory review and academic review have similar properties. Each has well-established criteria, although the weights given each will differ dramatically. Academics put a premium on elegance (although simplicity sometimes wins the day). Certainly, regulators will emphasize simplicity over complexity. Both seek robustness of results. The conclusions need to stand up to possible changes in initial conditions. Academics place a premium on the new and innovative, whereas regulators emphasize continuity. After all, telecommunications infrastructure assets are so long-lived that switching policies can wreak havoc with decision making. But there is also an element of continuity that academics do respect. The accepted paradigm will not be easily displaced, so the policy conclusions had better square with prior views with regards to the setting. It is okay for the results to be counterintuitive, so long as they are based on maximizing behavior by individual (generally, well-

TABLE 5.2
Telephone Service Quality Analysis: Lessons Learned From Serving Two Masters

1. The conceptual frameworks provided by economics and other decision sciences can shed light on extremely complicated phenomena.
2. The application of analytical and empirical tools to real-world problems requires a solid understanding of both the tools and the context in which they are being applied.
3. It is important to maintain close interactions with policymakers, checking to make sure that other stakeholders are not excluded from the process.
4. Present preliminary results in a variety of academic, corporate, and regulatory forums in order to benefit from expertise that exists outside your own research team.
5. Real policy change takes time—to obtain feedback, revise analyses, and convince decision makers of the merits of the change. Nothing is guaranteed.
6. Similarly, scientific review is a time-intensive process designed to (a) identify truly innovative and insightful ideas, and (b) screen out those that do not meet scholarly standards.
7. No one said it would be easy. If untenured, do not attempt to serve two masters!

informed) agents! Even more problematic for *Management Science* reviewers has been the team's perceived disregard for the marketing paradigm: "Quality is what customers perceive it to be." The team's efforts at explaining the role of experts in the regulatory process have (so far) been only partially successful.

Both processes are designed to kill bad or useless ideas. Thus, review lags are not only reasonable, but necessary if the contribution is to be evaluated carefully and thoughtfully. Rejections by an editor involve some randomness: The particular reviewer does not really understand the paper (alternatively, the points are not expressed in a logical and careful manner), the reviewer has an irrational grudge against a line of research (my favorite excuse), or the reviewer is on sabbatical and the disorganized editor lets one languish in purgatory for an unseemly amount of time. Whatever the reasons for the lag and rejection, the process is honored and is widely believed to improve the contributions to the scientific literature.

Good reviewers provide detailed feedback when the paper shows promise. Initial versions of papers are seldom ready for prime time. A parallel process occurs in the adversarial regulatory setting. Various stakeholders will identify limitations to the proposed policy. Alternatively, one group might stonewall an idea if implementation would be injurious to its position. The art of bargaining and compromise are probably better developed in the regulatory arena than in academia. The pursuit of "truth" can get in the way of "pretty good" policies.

So this tale still has no ending. The most comprehensive (and rigorous) expression of these ideas is still under review at a highly ranked journal. The feedback has been thorough and the team has tried to be responsive to reviewer suggestions. Similarly, the team still waits for Commission passage of a rule that just adds the proposed index to the regulatory toolkit for evaluating performance. In both cases, the lags seem long—but they are also understandable, given the stakes in both instances.

ADDENDUM

One week after the CITI Conference, the *Management Science* paper was accepted subject to minor revisions (see Lynch, Buzas, & Berg, 1994). The use of experts instead of consumers in developing weights no longer blocked academic acceptance of the conceptual framework. On June 14, 1993, the proposed FPSC rules became effective, so the index is now part of the regulatory process in Florida! The dual missions *were* accomplished.

ACKNOWLEDGMENTS

Helpful comments from John Lynch, Alan Taylor, Bill Lehr, and Jill Butler are acknowledged (without implicating them).

APPENDIX[12]

Hierarchical conjoint analysis circumvents problems with more commonly used variants of conjoint analysis: full profile analysis and two-at-a-time trade-offs. In the former, judges would have to keep information of all the scores in their minds—integrating them into an overall evaluation. Two-at-a-time trade-offs require the respondents to fill out matrices in which they evaluate profiles of combinations of high and low levels of performance of all possible pairs of dimensions. If there are 38 different dimensions, this involves filling out 703 different matrices.

The weighting scheme was developed by having experts from the FPSC and telcos rate different hypothetical company profiles of performance. Table 5.3 depicts one expert's ratings for four hypothetical combinations. Possible scores ranged from 1 (worst possible overall performance) to 10 (best possible overall performance). The observable range for high and low performance for each standard was established with the assistance of FPSC staff. Here, two aspects of service availability matter: (a) restoring primary service within three days of an outage, and (b) keeping appointments the morning or afternoon for which they are scheduled. Service restoration of 99% and meeting appointments 100% was rated 10 by this expert judge. If service restoration dropped to 88%, the combination was scored as an 8. Scores drop in a similar fashion if appointments are kept at the lower level, whereas restoration drops from 99% to 88%.

Thus, within the service availability cluster, one could distinguish between high and low levels of performance. When performance on the appointments standard is highest, the average score is $9((10+8)/2)$. Whereas when performance on the appointments standard is lowest, the average score is $5((6+4)/2)$. Thus, to calculate the weights within the service availability cluster, one can note the

TABLE 5.3
Rating Company Profiles

(a) Availability of Service

		Appointments = 100%	*Appointments = 94%*
3-Day Primary Service	99%	10	6
	88%	8	4

(b) Calculation of Effects on Ratings within Cluster

Dimensions	*Average for High Level*	*Average for Low Level*	*Difference Between Averages*
Appointments	(10 + 8)/2 = 9	(6 + 4)/2 = 5	4
3-Day Primary Service	(10 + 6)/2 = 8	(8 + 4)/2 = 6	2

(c) Calculation of Weights within Cluster

Dimensions	*Difference Between Averages*	*% Spanned*	*Weight Within Cluster*
Appointments	4	100 − 94 = 6	.667
3-Day Primary Service	2	99 − 88 = 11	.222

difference between these averages of 4. The number of percentage points spanned between high- and low-appointment performance is 4 performance points over the span of 6 percentage points. Thus, each 1 percentage point improvement in appointments implies a .667 increase in the score (on a scale of 1 to 10).

Note that the weights are the result of numbers assigned by evaluators (representing informed consumers). One expert respondent might assign a combination a 6 for a 3-day primary service and appointments of 99% and 94%, respectively. Another might assign a 7. The initial comparisons were ordinal in the sense that they are numerical representations of preference orderings. However, they have been treated as cardinal for policy purposes because a dollar metric was introduced in the second study that nailed the numbers to a dollar value. In addition, the numerical rankings across experts were statistically quite similar.

Table 5.4 illustrates the kinds of tests used to determine agreement among experts. The weights from four experts are shown. In this example, there is high disagreement regarding 3-day primary service. The sample mean (based on two observations from Experts 1 and 2) departs from the "true" weight that would be derived from all four experts. However, the data give a warning signal in that the sample standard deviation is high. For attributes in which there is low

TABLE 5.4
Weight of 1 Percentage Point Changes

Expert	*Appointments*	*3-Day Primary Service*
1	.200	.167
2	.220	.500
3	.180	.100
4	200	.400
Population Mean of 4	.200	.292
Population Standard Deviation	.014	.164
Mean Based on 1 & 2	.210	.334
Standard Deviation Based on 1 & 2	.014	.235

disagreement—appointments—the sample mean will be close to the "true" weight that would be derived from all four experts (here, the population). The low standard deviation captures this feature of the sample.

Similar respondent rankings for comparisons across clusters allowed the derivation of a comprehensive scoring formula containing weights for each 1-percentage point improvement in the different telephone service quality dimensions. The results allow one to have some confidence in the identification of quality dimensions that really matter and in the relative weights to be given those dimensions. There is always room for refinements, but the methodology represents a major improvement over pass–fail approaches now utilized by regulatory commissions.

ENDNOTES

1. Edwards (1977) found that when a series of multidimensional land-use proposals were evaluated intuitively and holistically by regulators and developers, the parties differed substantially in their rank orders of the proposals' desirability. When the relative values the parties placed on the dimensions were measured and modeled, the evaluations derived from the models of all parties exhibited remarkable agreement. Edwards suggested that the disagreements in holistic judgments stem from unconscious tendencies to focus on the subset of dimensional cues that are consistent with the overall judgment the parties would like to draw. In the present context, such selective processing can lead company representatives to truly believe that their companies provide superior quality, whereas skeptical regulators reach a contrary conclusion from the same evidence.
2. Note that the scoring function presented earlier is merely a transformation of the constraint equation described here:

$$
\begin{aligned}
Q_a &= W_0 + W_1(Z_1 - Z_1^*) + W_2(Z_2 - Z_2^*) \\
Q_a &- (W_0 - W_1Z_1^* - W_2Z_2^*) = W_1Z_1 + W_2Z_2 \\
Q &= W_1Z_1 + W_2Z_2
\end{aligned}
$$

3. Whitley (1991) argues that "The role of journals in economics is closely connected to the nature of economics as a scientific field . . . journals dominate the formal communication system,

are ordered into a strong prestige hierarchy, reproduce a strong analytical orthodoxy and publish highly formal, standardized and concise papers" (p. 32).

4. Given the difficulties in sorting out potential cost complementarities, the quality dimensions might be considered in related bundles.
5. Noam (1991) addressed numerous implementation issues, including the integration of service quality indices into the incentive process. For example, "Gold-plating could . . . be dealt with by setting ceiling for rewards" (p. 185).
6. The use of 75 as the base score for just meeting each standard involved a simple transformation of the calculated coefficients, although it complicates comparisons with the initial study (of 38 dimensions).
7. An alternative, methodologically cleaner, approach would be to raise the passing score to 80 or 85, in which 75 represented just meeting all initial standards.
8. As has already been noted, the differential weight for sub- and superperformance was a feature of the second study that might be capturing a regulatory concern for goldplating.
9. The derivation is available in Lynch et al. (1994). Note that the proposed rule utilizes weights for a 100-point scale, in which 75 is the score for just satisfying each rule.
10. I am indebted to John Lynch for formulating the points related to customer versus expert perceptions.
11. A similar approach was recommended to the New York Public Service Commission in a Theodore Barry & Associate (TB&A) study on performance improvement opportunities at the New York Telephone (NYT) Company (Mayer, 1993). Their service quality index had seven elements: dial line, customer contact, maintenance, installation, customer expectation, and operational efficiency. The TB&A study evaluated NYT programs in terms of impacts on these categories. It attempted to identify trade-offs among construction, maintenance expense levels, and service quality. As such, the approach provides both a management tool (for investment planning) and a regulatory analytic technique (for evaluating corporate performance). The TB&A line of investigation illustrates the increased attention being given to indexes and weights for addressing telephone service quality issues.
12. Thanks to Tom Buzas for developing the example used in the Appendix and in Table 5.3, and to John Lynch for the example in Table 5.4.

REFERENCES

Berg, Sanford V., & John G. Lynch, Jr. (1992). The measurement and encouragement of telephone service quality. *Telecommunications Policy, (April)*, 210–224.

Buzas, Thomas E., & John G. Lynch, Jr. (1988). *A formula for the comprehensive evaluation of local telephone companies: Report to the Florida Public Service Commission: Preliminary report.* (January).

Buzas, Thomas E., John G. Lynch, Jr., & Sanford V. Berg (1991). Issues in the measurement of telephone service quality. In Barry Cole (Ed.), *After the breakup: Assessing the new post AT&T divestiture era* (pp. 268–276). New York: Columbia University Press.

Edwards, E. (1977). How to use multiattribute utility measurement for social decision making. *IEEE Transactions in Systems, Man and Cybernetics, 7*, 326–340.

Feldman, J. M., & John G. Lynch, Jr. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73*, 421–435.

Fischhoff, Baruch (1993). Value elicitation: Is there anything in there? In M. Hechter, L. Nadel and R. Michod (Eds.), *The Origin of Values* (pp. 187–214). Hawthorne, NY: Aldine de Gruyter.

Louviere, Jordan (1984). Hierarchical information integration: A new method for the design and the analysis of complex multiattribute judgement problems. In Thomas Kinnear (Ed.), *Advances in consumer research, Vol. 11* (pp. 148–155). Provo, UT: Association for Consumer Research.

Lynch, John G. Jr., Thomas E. Buzas, & Sanford V. Berg (1994). Regulatory measurement and evaluation of telephone service quality. *Management Science, February*, 169–194.

Mayer, Robert H. (1993). *Integrating service quality, customer service and alternative investment analysis.* TB&A Management Consultants, April 23.

Noam, Eli M. (1991). The quality of regulation in regulating quality: A proposal for an integrated incentive approach to telephone service performance. In M. A. Einhorn (Ed.), *Price caps and incentive regulation in telecommunications* (pp. 167–189). Boston: Kluwer Academic Publishers.

Rovizzi, Laura, & David Thompson (1992). The regulation of product quality in the public utilities and the citizen's charter. *Fiscal Studies, 13:3*, 74–95.

Whitley, Richard (1991). The organization and role of journals in economics and other fields. *Economic Notes, 20:1*, 6–32.