Optimal Peak-Load Pricing for Local Telephone Calls

Rolla Edward Park, Bridger M. Mitchell
The preparation of this report was supported by The RAND Corporation as part of its program of public service.


The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of The RAND Corporation do not necessarily reflect the opinions or policies of the sponsors of RAND research.

Copyright © 1987
The RAND Corporation

Published by The RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138
Optimal Peak-Load Pricing for Local Telephone Calls

Rolla Edward Park, Bridger M. Mitchell

March 1987

RAND
PREFACE

This report explains how to calculate optimal time-of-day rates for local telephone calls. To make the illustration concrete and realistic, the authors use detailed telephone traffic data for a small city in Illinois in a computer model that simulates the welfare effects of alternative tariffs.

Some strong conclusions emerge from the analysis. One conclusion concerning the economic efficiency of measured service prices for local telephone calls should interest anyone who seriously cares about local telephone prices, including public utility commissioners and their staffs, telephone company executives, and consumers of telephone services. Another conclusion concerning the application of peak-load pricing theory to other practical problems should interest economic theorists and analysts who deal with peak-load pricing theory or its application.

This report is one product of a project that began in 1977 with a major grant from the National Science Foundation. A second grant from NSF, together with one from General Telephone and Electronics, supported the development of the model used in this report. From its own funds, The RAND Corporation supported the applications of the model reported here, as well as the preparation of this final report. Other publications of the larger telephone pricing project are listed within (pp. ix-x).


Five technical appendixes are separately bound in a companion volume, R-3404/1-1-RC.

Correction Note. The first edition of this report (June 1986) was based on an erroneous data file. This second edition corrects the error.

We made available to Bell Communications Research our computer simulation program (Appendix D) and (with the approval of GTE Service Corporation) its GTE-traffic-data input files. While working with our program, Martin Koschat, Dale Lehman, and Elaine Sieff at Bellcore discovered an error in the traffic data for business customers.
Checking back, we traced the error to the time when we created the business input file. The month of April was missing from the raw business data. We attempted to fill in each missing hour by averaging the traffic during two similar hours: one hour was the corresponding hour-of-day and day-of-week exactly five weeks earlier than each missing hour and one hour was the corresponding hour five weeks after each missing hour. We erred and summed the traffic instead of averaging it. Thus, the business data for April were overstated by a factor of two. The effect of the error was to create a sharp false busy season in April, greatly exaggerating the seasonal variance in calling rates.

We have corrected the business data and recomputed all of the numbers that are affected by the correction. This corrected version of our report reflects the new results. The numbers have changed, some of them substantially, but we believe that our former conclusions are unaffected. Variation in demand for telephone calls within feasible pricing periods sharply limits the efficiency gains that price rationing can achieve. Given the expected low levels of incremental capacity costs for modern switching and transmission technologies, measured-rate pricing of local telephone calls is likely to be slightly less efficient than traditional flat-rate pricing.

In addition to correcting the data error, we have added a brief "Postscript" section, hoping to clarify a number of points raised by readers of the original version of the report.

\[1\text{We also had to fill in smaller periods of missing data in both the business and residence traffic files. We managed to do these calculations correctly.}\]
SUMMARY

This report shows how to find optimal time-of-day measured-rate prices for local telephone calls. We use a simulation model based on actual telephone traffic data for each hour for a full year. The model calculates capacity cost savings, measurement costs, losses in consumer benefit due to price rationing, and losses due to quantity rationing, to assess the net welfare effects of alternative tariffs.

This is the first application of peak-load pricing theory to recognize and account for variation in demand within feasible pricing periods. Feasible tariffs are limited to perhaps three price periods that repeat from day to day, and local telephone demand varies markedly within such periods. This variation sharply limits the efficiency gains that price rationing can achieve, because feasible prices are inevitably too high during some hours when more calls could be served at no incremental cost, and too low during others when serving more calls requires large expenditures to increase capacity.

We find that, contrary to conventional wisdom, measured-rate pricing of local telephone calls is likely to be somewhat less efficient than traditional flat-rate pricing. If local measured service is desirable public policy, it must be justified on grounds other than economic efficiency.
ACKNOWLEDGMENTS

In the course of this research, we have benefited from continuing discussions with Gerald Cohen and Philip J. Crabill of GTE Service Corporation. Ronald R. Braeutigam of Northwestern University and RAND colleague Daniel F. Kohler suggested improvements to the final draft.

We also appreciate the helpful comments and suggestions of James H. Alleman (International Telecommunication Union), Edward C. Beauvais (GTE Service Corporation), Lawrence P. Cole (GTE Service Corporation), John M. Hartwick (Queen's University), John M. Jensik (General Telephone Company of California), Lawrence I. Little (General Telephone Company of California), Willard G. Manning (The RAND Corporation), Herman Quirmbach (University of Southern California), Judith Wilson Ross (University of California, Los Angeles), Ray Schultz (General Telephone Company of California), Neal Stollem (GTE Service Corporation), E. Subissati (Bell Canada), Helen B. Turin (The RAND Corporation), and participants in the Tenth and Twelfth Telecommunications Policy Research Conferences, the Fourth International Forecasting Conference, and seminars at The RAND Corporation and at Bell Communications Research.

In addition, we are grateful to Martin Koschat, Dale Lehman, and Elaine Sieff of Bell Communications Research for uncovering an error in one of the input data files used in the first edition of this report. The Bellcore group, Edward C. Beauvais of GTE Service Corporation, William B. Shew of National Economic Research Associates, Lester D. Taylor of the University of Arizona, and Jan Paul Acton of The RAND Corporation all shared with us their detailed critiques of the report.
TELEPHONE PRICING PROJECT
PUBLICATIONS


Park, R. E., B. M. Mitchell, and B. M. Wetzel, "Demographic Effects of Local Calling Under Measured vs. Flat Rate Service: Analysis of Data from the GTE Illinois Experiment," in Pacific


CONTENTS

PREFACE ........................................ iii
SUMMARY ....................................... v
ACKNOWLEDGMENTS .............................. vii
TELEPHONE PRICING PROJECT PUBLICATIONS ...... ix
FIGURES AND TABLES ........................... xiii

Section
I. INTRODUCTION ............................... 1
   Applied Peak-Load Pricing Theory ............ 1
   Public Policy Toward Telephone Pricing ...... 2
   Some Limitations ........................... 3
   Overview .................................... 4

II. THEORY .................................... 5
   Pure Pricing Solutions ...................... 5
   Pure Quantity-Rationing Solutions ......... 6
   Combined Price and Quantity Rationing ..... 7
   A Graphical Example ....................... 7

III. DATA ..................................... 10

IV. MODEL .................................... 14
   Capacity Costs ............................. 14
   Measurement Costs ......................... 17
   Price Rationing ............................ 17
   Quantity Rationing ......................... 20
   Using the Model ............................ 23

V. RESULTS .................................. 24
   Base-Case Assumptions for Technology in Common
   Use Today .................................... 24
   Sensitivity Analyses ....................... 30

VI. CONCLUSION ............................... 32
   Applied Peak-Load Pricing Theory .......... 32
   Public Policy Toward Telephone Pricing .... 34
Appendix¹

A. DOES CONSTRUCTION OF ESTIMATED CALLING RATES FOR THE TOTAL RESIDENTIAL POPULATION FROM A ONE-SIXTH SAMPLE OVERSTATE THE VARIABILITY OF RESIDENTIAL CALLING?

B. SIMULATION OF HOURLY DEMAND CROSS-ELASTICITY FOR LOCAL TELEPHONE CALLS

C. CALCULATING CHANGES IN CONSUMER BENEFITS

D. COMPUTER PROGRAM

E. DETAILED TABLES OF RESULTS

¹The appendixes are separately bound in a companion volume: Optimal Peak-Load Pricing for Local Telephone Calls: Technical Appendixes, R-3404/1-1-RC.
FIGURES

1. Graphical Example of Price and Quantity Rationing ........ 8
2. Average Busy-Season Hourly Calls ....................... 12
3. Load-Duration Curve of Calls ............................. 13

TABLES

1. Annualized Usage Costs of Capacity from the Arizona
   Cost Study ................................................... 15
2. Own-Price Coefficients in Calibrated Demand Model ...... 19
3. Welfare Effects of Local Measured Telephone Service
   in a Small Exchange: Base-Case Assumptions for
   Technology in Common Use Today .......................... 25
4. Sensitivity Analysis Cases .................................. 30
5. Sensitivity Analysis Welfare Results ........................ 31
6. Summary of Base-Case Results for Technology in
   Common Use Today ......................................... 33
I. INTRODUCTION

Local telephone rates are likely to change substantially in the next few years. Rates have been kept low by a cross subsidy from long distance service (Rohlfs, 1979; Meyer et al., 1979). Now, however, with the breakup of AT&T and increasing competition in the long-distance market, the subsidy is being phased out and local rates are rising.

Residential telephone service in most of the United States has been offered for a fixed monthly subscription fee, with no extra charge for local calls, a pricing method that is commonly referred to as “flat rates.” Telephone companies are now trying new pricing schemes to ease the transition to higher local rates. Notable among the new pricing methods are tariffs that charge separately for access to the telephone network and for telephone use; such tariffs are called “measured rates” or “local measured service.”

This report shows how to find optimal time-of-day measured-rate prices for local telephone calls. We use a simulation model based on detailed telephone traffic data to calculate the welfare effects of alternative tariffs. Our study has important implications for other applications of peak-load pricing theory as well as for public policy toward telephone pricing.

APPLIED PEAK-LOAD PRICING THEORY

Almost all of the costs of supplying local telephone service are capacity costs; the marginal cost of a call during offpeak hours is very close to zero. Surprisingly, however, neither previous analyses of telephone pricing (Alleman, 1977; Mitchell, 1978) nor the best known papers on peak-load pricing (Boiteux, 1949, 1960; Steiner, 1957; many others) offer much guidance about how to design optimal time-of-day tariffs for telephone calls.

Both Alleman and Mitchell largely ignored the peak-load problem. Instead of examining capacity costs, they simply assumed that there was a constant marginal cost per telephone call that could be used to calculate efficient prices and did not investigate the source of that cost.

1The electric utility industry uses “flat rates” differently, to distinguish between prices that are uniform around the clock and time-of-day prices. Readers accustomed to that usage should be on their guard against the potential confusion.

2Another common name for this pricing method is “usage sensitive pricing.”
Littlechild and Rousseau (1975) analyze fixed peak, shoulder, and off-peak pricing periods for local and long-distance calls using intra-peak peak-demand factors. However, they do not investigate the welfare effects of changing the pricing periods or of rationing capacity. And most of the general literature on peak-load pricing theory also fails to consider the effect of intra-period demand variation.³

We believe that our study is the first application of peak-load pricing theory to account for the efficiency effects of variations in demand within feasible pricing periods.⁴ Feasible tariffs are limited to perhaps three price periods that repeat from day to day, and local telephone demand varies markedly within such periods. This variation sharply limits the efficiency gains that price rationing can achieve, because feasible prices are inevitably too high during some hours when more calls could be served at no incremental cost, and too low during others when more calls require large expenditures on added capacity.

Intra-period demand variation is likely to be important in other markets with fluctuating peak loads, such as electric power. Studies that design or evaluate time-of-day prices for such markets should account for demand fluctuations within pricing periods. Of course, the conclusions of such studies may be quite different from our results for local telephone service, because they will depend on demand and costs in each industry. The key point is that intra-period demand variation has the potential to affect the conclusions substantially and should not be ignored.

PUBLIC POLICY TOWARD TELEPHONE PRICING

Since Mitchell (1978), most economists and informed communications policymakers have probably assumed that measured-rate prices for local telephone calls could substantially increase economic efficiency, at least if the additional costs for measuring and billing were low enough.⁵ Over the past several years, state regulatory commissions have approved various measured-rate pricing plans in several

³An elegant but largely overlooked paper by Danaby (1975) is an exception. Danaby provides a complete theoretical solution to the rate design and capacity choice problem with intra-period varying demand. In addition, Craven (1971, 1985) establishes the optimal peak and off-peak periods for a special case that requires capacity to be sufficient to serve the maximum demand.

⁴In this context, feasible means simple enough for the consumer to grasp.

⁵Those who have opposed local measured service on efficiency grounds (for example Selwyn, 1982) have stressed factors that decrease the efficiency of measured rates in Mitchell’s model—low demand elasticity, low marginal costs per call, high measuring costs.
jurisdictions, and telephone companies' proposals are pending in many others.

We find that, contrary to conventional wisdom, measured rates will at best produce very modest efficiency gains; more likely, they will result in small efficiency losses. Any efficiency difference between flat rates and measured rates is probably not large enough to dictate the choice between the two pricing methods. Other considerations, in particular the distributional and competitive effects of the two methods, should be carefully weighed before the choice is made.

SOME LIMITATIONS

This report examines economic efficiency in the market for local telephone calls. We neglect effects in other markets, as well as the effects on policy goals other than efficiency, to reach our central results, which show how intra-period demand variation can limit the gains from price rationing.

Effects in Related Markets

In principle, changing the prices charged for telephone calls would affect consumers' demand for access to the telephone network (Allenman, 1977; Mitchell, 1978); if measured rates are substituted for flat rates without lowering the monthly charge, some households may cancel their telephone service. However, if the extra revenues from measured rates are used to lower the access price, the number of subscribers might increase. Changes in the number of subscribers could have two types of welfare effects: a change in direct consumer benefits and a change in subscriber externalities. We assume that the access demand elasticity and the cross-elasticity between calls and access are small. Consequently, the welfare effects of any changes in access prices are also small.\(^6\)

Long-distance carriers as well as consumers need access to the local telephone switch. This service is now supplied under usage-sensitive

---

\(^6\)The available empirical evidence indicates that they are (Taylor, 1980; Perl, 1984).

\(^7\)The demand for access lines may become more elastic in the future if "shared-tenancy" arrangements proliferate. In a typical arrangement, a building owner installs a private switch to handle the occupants' telephone calls and rents only enough access lines from the telephone company to accommodate their combined peak traffic. Clearly, the price of access lines affects the incentives to provide such service. By decreasing the price of access to tenants of multi-tenant buildings but raising it to single-line customers, shared-tenancy arrangements might either decrease or increase the total number of telephone subscribers.
tariffs but could just as well be supplied at flat rates. We have not studied the efficient pricing of carrier access or the efficient pricing of the long-distance calls themselves.

Policy Goals Other than Economic Efficiency

The choice between measured rates and flat rates will affect nearly everyone's telephone bill and may also change the number of telephone subscribers, effects that cannot be omitted from public policy deliberations. Some households will pay higher bills and others will pay less if flat rates are changed to measured rates. These distributional effects may be of particular concern if certain groups (the elderly, the poor) are differentially affected. And if measured rates affect the number of telephone subscribers, they are of concern to the venerable policy goal of "universal" telephone service. Measured rates may also help local telephone companies to compete for customers who would otherwise bypass the local switched network or enter shared-tenancy arrangements to reduce their need for access lines (Brock, 1986).

This report concentrates on economic efficiency, leaving other public policy goals for other studies.

OVERVIEW

We review peak-load pricing theory in the next section, paying particular attention to the usually overlooked role of intra-period demand variation. In Sec. III, we describe the very detailed local telephone traffic data that we use. Section IV outlines how our model calculates the four major components of change in welfare that might result from a change in local telephone pricing: capacity cost savings, measuring and billing costs, losses in consumer benefit due to price rationing or to quantity rationing. Results of the welfare calculations for various tariffs of both theoretical and policy interest are in Sec. V. Section VI contains the concluding discussion.
II. THEORY

PURE PRICING SOLUTIONS

Telephone service looks like the classic case of a service suitable for peak-load pricing. With minor exceptions, the variable costs of an automatic telephone network are a function of the network's maximum capacity (or else they vary only with the number of subscribers). At hours when spare capacity is available, serving an additional call requires almost no additional resources. But when capacity is fully used, additional calls cannot be completed, and the cost of supplying an extra call is the incremental cost of expanding network capacity.  

The two-period Boiteux (1949, 1960)-Steiner (1957) model is the simplest stylization of these facts. In that model, demand is at a uniform level in each of two periods and capacity must be sufficient to satisfy the maximum demand. The optimal prices are a peak price equal to the marginal cost of expanding capacity, and an offpeak price equal to zero. Optimal capacity is equal to the peak demand at these prices.

Applied mechanically to local telephone service, a trial solution would set a peak price of perhaps $100 per call for the peak hour of the year. Clearly, such a high price would radically depress demand in that hour far below the second highest hour, thus causing a "shifting peak." In that case, the optimal solution requires pricing several hours at nonzero rates (Boiteux, 1949, 1960).

If a separate price can be set for each high use hour, the optimal first-best solution is a set of prices that cause the load in each high use hour to be reduced to a common level, equal to the capacity of the network. A sufficient number of hours must be priced that the sum of all prices equals the marginal capacity cost. For constant marginal capacity cost, total revenues will then equal total costs.

In practice, this solution is not feasible for at least two reasons. First, one does not know in advance the exact timing of all of the high

---

1In reality, calls arrive randomly at the telephone central office, and the greater is the traffic at the moment they arrive the longer the delays and the higher the probabilities of not being served. We approximate what is really a sharply curved congestion function with a right-angled one. Also, we take one hour as the basic unit of time and assume that demand is uniform within the hour. We make these assumptions so that we can concentrate on the effects of fluctuations in hourly traffic; accounting for intra-hourly fluctuations as well would greatly complicate our analysis and certainly strengthen our conclusions. For an explicit treatment of the randomness of telephone calling, see Kay (1979).
use hours and the loads that will be offered in each. Second, even if perfect information enabled one to calculate the first-best price schedule, the tariff would be much too complicated for telephone subscribers to cope with. Perhaps computer technology will someday support a scheme in which prices based on instantaneous system loads are quoted to consumers when they place their calls, thus overcoming both problems. But for the present, feasible time-of-day prices must be set in advance for a small number of periods, such as regular weekday hours.

Welfare gains from feasible prices will thus fall short of the gains attainable by the complicated first-best set of prices. Feasible prices will not succeed in leveling the load during all high use hours to equal available capacity and, inevitably, they will apply unnecessarily during many hours when loads would fall short of capacity even at a zero price.

**PURE QUANTITY-RATIONING SOLUTIONS**

To date, time-of-day pricing of local telephone calls is the exception, not the rule; telephone companies currently rely on quantity rationing to limit demand to available capacity. To serve the absolute peak demand that occurs during only one or a very few hours of the year by building expensive capacity that is idle all the rest of the year is clearly uneconomical. Therefore, telephone companies follow capacity engineering rules intended to meet some specified blockage probability during design busy hours (.01 is commonly used), or to serve some specified average load rather than the peak demand.

Economists frequently assume that price rationing must be more efficient than quantity rationing, but that need not be true for telephone calls. To be sure, price rationing excludes only the lowest valued calls, while quantity rationing indiscriminately excludes both high and low valued calls. But feasible (two- or three-period) peak-load prices apply indiscriminately to all hours within broad pricing periods, while quantity rationing excludes calls only during the scattered hours with the highest calling rates.

The rule for determining the optimal level of capacity under quantity rationing is similar to the rule for determining optimal hour-by-hour first-best prices: capacity should be chosen so that the sum of the values of serving a marginal call during each of the rationed hours equals the cost of a marginal unit of capacity. In fact, if somehow calls could be rationed in order of their value (so as to eliminate only the lowest valued calls), then the allocation that would result from optimal
quantity rationing would be identical to the allocation that results from optimal price rationing.

Quantity rationing does not, however, ordinarily provide a mechanism for singling out low valued calls. If rationing is random, as is commonly assumed, calls not served will tend to be of average value. Since unserved (lost) calls are on average worth more than they are under price rationing, the optimal capacity will be greater than with optimal price rationing.²

COMBINED PRICE AND QUANTITY RATIONING

We are accustomed to thinking of price rationing and quantity rationing as alternative methods for allocating resources. Policy-oriented researchers sometimes explore the relative merits of these allocative methods in particular applications, such as emissions control (Palmer et al., 1980) or airplane runway delays at busy airports (Carlin and Park, 1970).

In the present case, price rationing and quantity rationing may be complements (Dansby, 1975). Price rationing excludes only the lowest valued calls, whereas quantity rationing excludes both low valued and high valued calls. Quantity rationing, however, excludes calls only during those hours when demand (offered load) exceeds capacity, whereas any feasible price rationing scheme will exclude calls during some hours when they could be served at no incremental cost.

A GRAPHICAL EXAMPLE

Figure 1 illustrates the effect of price and quantity rationing when there is intra-period demand variation.³ The illustration uses a graph of the number of unrationed calls made during the ten hours of an imaginary (short) day. The hours are arranged and labeled in decreasing order of the number of calls made when demand is totally unrationed. In panel A of Fig. 1, hourly calls range from 1000 during hour 1 down to 100 during hour 10. This ordering of hours produces what the electric power industry calls a “load-duration curve.” Hours that

²Random rationing may or may not be a reasonable assumption for telephone calls. The value of serving an additional call that otherwise would be lost to quantity rationing may be even greater than the value implied by the random rationing assumption. In this report, we use a higher value implied by a stylized version of the capacity decision rules now used by telephone companies (see Sec. IV).

³This is not meant to be any more than an heuristic introduction to the choice among tariffs; a complete welfare comparison of alternative tariffs occupies the entire remainder of the report.
Fig. 1 — Graphical example of price and quantity rationing
are adjacent on a load-duration curve are not necessarily adjacent in
time: in the figure, the white bars represent daytime hours and the
crosshatched bars nighttime hours. Although the nighttime hours tend
to have fewer telephone calls than the daytime hours, and thus to
appear mostly toward the right side of the load-duration curve, this is
not true for all nighttime hours; in this example, one especially busy
nighttime hour has the third highest number of calls.

*First-best peak-load pricing* establishes different prices for each peak
hour to level the load at the optimal capacity, shown arbitrarily as 500
calls in panel B of the figure. This requires a high price during hour 1
and successively lower prices during hours 2 through 5.\(^4\) Just the lowest
valued calls during each priced hour are eliminated by price rationing;
these rationed calls are shown stippled in the figure.

Separate prices for each hour may not be feasible (especially if we
are dealing with some 8700 hours of the year, instead of just 10 hours
as in our example). Panel C shows what happens if the tariff is limited
to *two feasible prices*—one daytime and one nighttime price. The unif-
form daytime price reduces calling during all daytime hours, whether or
not they press on capacity.\(^5\) With pure price rationing, capacity must
be built to serve the maximum number of calls per hour, shown as 800
in the figure. Uniform prices eliminate some calls that could be served
costlessly during hours 2, 4, 5, and 6.

*Quantity rationing* may be an attractive alternative to feasible price
rationing. In panel D, quantity rationing requires capacity equal to the
fourth highest hourly demand (shown as 700 calls) and results in
excess calls (diagonal lines) during the three highest hours. The unserved
calls are, on average, worth more than the lowest valued calls
that would be eliminated by price rationing; but there is no unnesses-
ary loss during offpeak hours, because all of the demand during less
busy hours is served.

Finally, it may be advantageous to *combine price and quantity
rationing*, as in panel E. Here a lower daytime price than in panel C
and a smaller amount of quantity rationing than in panel D balance
the marginal losses due to each rationing method. This combined
approach may be more efficient than either price or quantity rationing
alone.

\(^4\)To keep the example simple, we assume that the demand curves for all hours are
linear, parallel, and independent (no cross-elasticity).

\(^5\)We assume here that the optimal nighttime price is zero.
III. DATA

General Telephone and Electronics conducted an experiment in three small cities in Illinois to determine the effects of charging for local telephone calls.\textsuperscript{1} We use data on local residential and business telephone calls in one of the GTE experimental exchanges—Clinton, Ill.—during 1976, when flat monthly rates (with no extra charge for calls) were in effect. Starting in May 1975, and continuing to the present, GTE has recorded information on each local call made in the three experimental exchanges, including calling number, time of call, and duration of call.

We processed these call-detail records for 476 single party residential telephone numbers to create a file of residential telephone use by hour of day throughout the year. The file includes both number of calls and minutes of calling during each of 8784 hours of the year; in this study, we use the data for calls, but not minutes, and we omit the last two days of the year to get an even 52 weeks (8736 hours). The 476 telephone numbers are all of those that were sampled in the first three waves of a survey conducted for GTE by Opinion Research (1980), except for 70 that we randomly excluded to undo the effects of the oversampling of unusually heavy and unusually light telephone users in the third survey wave. We inflated the residential hourly-use file by the factor 3035/476 to approximate use by all 3035 single party residential subscribers in Clinton in mid-1976.\textsuperscript{2}

The business hourly-use file is based on call detail records for all 575 business, key telephone, and private branch exchange lines that were in continuous subscription in Clinton during 1976. For both business and residential calls, there were some gaps in the call detail records, which we filled in by averaging calls during matching hours and days of the week before and the week after the gap.

The pattern of telephone calling on the average weekday of the “busy season” is plotted in Fig. 2. The busy season is defined by GTE as the four consecutive weeks of the year with the highest total number

\textsuperscript{1}See Cohen (1977) for a description of the study, and Park, Wetzel, and Mitchell (1983) for estimates of demand elasticities.

\textsuperscript{2}In principle, this technique will overestimate the hour-to-hour variability in the number of calls unless individual households’ calling patterns are perfectly correlated over time. In practice, we concluded that the overestimate is not large enough to affect our results very much. See Appendix A.
of calls. Telephone use peaks in the morning from 0900 to 1000 (hour 10 on the chart) and again in the afternoon from 1600 to 1700 (hour 17). Business use constitutes 37 percent of all calling during business hours (0800 to 1700). As discussed below (Sec. IV), busy-season data or similar data are the basis for capacity planning for most telephone companies.

Figure 3 shows the "load-duration" curve for calls. To plot a load-duration curve, we rank each hour of the year according to the number of calls placed during that hour, then plot calls against rank. The spike at the left side of Fig. 3 indicates that an exceptional number of calls are placed during a few hours of the year. The very busiest hour plotted had 1968 calls. Because flat rates were in effect, there was no price rationing in Clinton during 1976. We assume that there was no quantity rationing either, so that the observed loads are equal to the quantity demanded at a zero price. The spikiness of the load-duration curve supports this assumption.

---

3In this case it is November 29 through December 26.
Fig. 2—Average busy-season hourly calls
IV. MODEL

Using the Clinton 1976 calling data described above, our model evaluates the welfare effects of alternative local measured service price and capacity decisions. The model calculates four components of welfare: capacity cost savings, measurement and billing costs, price rationing losses, and quantity rationing losses. In this section, we describe how we established the base-case values for each of the four components.

We also describe the variations on base-case values, used in sensitivity runs, that together represent demand and cost conditions in a wide variety of local telephone markets. The base-case cost values are drawn from publicly available studies of switching and transmission technologies that have been widely used in the past. Some unpublished industry studies suggest that modern digital switching and fiber optic transmission facilities—the technologies that will predominate in current and future installations—have substantially lower incremental capacity costs. The costs of these modern technologies are probably better approximated in the cases with lower capacity and measuring costs than in the base case.

CAPACITY COSTS

Capacity costs will vary with the number of calls (or minutes) that the telephone switching equipment and trunk lines are designed to serve. We are interested in the additional cost of serving one additional call per hour\(^1\) and use the annual equivalent of the incremental investment cost, reflecting such factors as yearly interest, depreciation, maintenance, and taxes.

For our base case, we assume that the annualized incremental cost of capacity is $100 per call per hour. For example, this assumption means that the cost of plant capable of serving 1501 calls per hour is $100 per year more than the cost of plant with a capacity of 1500 calls per hour. This is a stylized cost figure for local calls that use only one switching office, based on Mountain Bell's (1980) measured service filing for much of Arizona.

\(^1\)Because the equipment is available only in “lumps,” the cost must actually be estimated as the average incremental cost over some wider range of alternative capacities.
Table 1 shows the annualized usage costs of capacity from the Arizona cost study,\(^2\) excluding measurement and billing costs, which are discussed below. We do not know what annualization factors Mountain Bell used; we do know that they “account for depreciation expense, cost of money, income taxes, administrative expense, maintenance expenses, ad valorem tax, gross receipts tax, and license contract fees.”\(^3\) The costs are expressed in 1984 dollars for the switching and transmission technologies that were expected to be used in Arizona in 1983. Apparently those technologies did not include much digital switching or fiberoptic transmission. They did include a substantial fraction (15 to 20 percent) of older “crossbar” switches.

**Sensitivity Analysis**

In sensitivity analyses, we use a range of capacity costs, both higher and lower than the base-case value of $100 per call. The higher value, $200 per call, is meant to represent an average cost for areas with both

<table>
<thead>
<tr>
<th>Distance</th>
<th>Call Setup</th>
<th>Per Minute</th>
<th>Total for a 3.5 Minute Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-switch</td>
<td>33</td>
<td>19</td>
<td>100</td>
</tr>
<tr>
<td>Inter-switch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 6 miles</td>
<td>90</td>
<td>34</td>
<td>210</td>
</tr>
<tr>
<td>6-10 miles</td>
<td>90</td>
<td>36</td>
<td>215</td>
</tr>
<tr>
<td>10-16 miles</td>
<td>91</td>
<td>38</td>
<td>223</td>
</tr>
<tr>
<td>16-22 miles</td>
<td>96</td>
<td>52</td>
<td>278</td>
</tr>
<tr>
<td>22-30 miles</td>
<td>101</td>
<td>66</td>
<td>331</td>
</tr>
<tr>
<td>30-40 miles</td>
<td>103</td>
<td>72</td>
<td>354</td>
</tr>
<tr>
<td>40-55 miles</td>
<td>109</td>
<td>87</td>
<td>413</td>
</tr>
</tbody>
</table>

\(^2\)We made minor adjustments to the cost figures in Mountain Bell (1980, exhibit 14, pp. 17–19). Specifically, we excluded measurement costs (which are treated separately below), translated from “first minute/additional minute” to “setup/minutes,” and inflated to 1984 dollars using the ratio of the Handy-Whitman communications cost index for 1984 and 1983 (215/208).

\(^3\)Mountain Bell (1980, p. 6).
intra-switch and inter-switch local telephone service, using a mixture of technologies that are already in place.\textsuperscript{4} Some proprietary telephone company studies suggest that incremental capacity costs for digital switching and fiber optic transmission are substantially lower than the costs for older technologies. We use capacity costs of $50 and $20 per call to represent possible costs of digital and fiber optic technologies, which will be the predominant ones to be installed in the near future.

**Engineered Capacity**

The combination of costly capacity and a spikey load-duration curve means that it is not optimal to build a telephone plant to accommodate the maximum number of calls per hour ever offered. If 1976 were the design load year for Clinton, it would cost $100 to serve the last (1968th) call during the absolute peak hour (Fig. 3). That call would have to be valuable indeed to justify such a high cost.

Telephone company engineering practice recognizes this fact by choosing capacity to serve less than the absolute peak number of calls expected during the design year. The capacity rule may be expressed in terms of a specified blockage probability (.01 is commonly used), or by designing for some specified average load rather than the peak load. We shall make use of the following stylized statement of current engineering practice.

*Engineering capacity rule:* Design capacity is determined by first finding the busiest consecutive four-week period of the year, then finding the busiest hour of each of the 20 weekdays during that period, then choosing capacity so that all except 1 percent of the calls placed during those 20 hours can be served.

During 1976, Clinton had a total of 27,152 calls during the 20 busy-season busy hours. Thus the engineering rule chooses capacity so that 272 (1 percent) of them are not served. A capacity choice of 1533 calls per hour excludes a total of 272 calls during three of the busy-season busy hours. (It is sufficient to handle all of the calls placed during the other 17 hours.) In addition, that capacity choice excludes a total of 1392 calls during eight other hours of the year (including the absolute peak hour, which did not in fact occur during the busy season).

The engineering rule is not the result of explicit capacity optimization (although we show below how it can be interpreted as implicit optimization). Our model normally chooses capacity according to the

\textsuperscript{4}About half of all local calls in the United States are intra-switch calls (Mitchell, 1983).
engineering capacity rule, to reflect actual telephone company practice. In some sensitivity analyses, we also explore explicit capacity optimization.

MEASUREMENT COSTS

Beauvais (1984) comprehensively discusses the costs of measuring and billing local calls. For one particular company (General Telephone of Wisconsin), he calculates an investment cost of $10.30 per line for measuring and billing equipment, and an ongoing incremental cost of $0.00104 per call for measurement and bill processing. Applying a 20 percent annualization factor to Beauvais’s figures, we adopt base-case values of $2.10 per line and $0.001 per call. These are the costs to produce routinely a bill with a summary of total local calls and minutes by distance band and time of day.5

Sensitivity Analysis

Measurement and billing costs will be higher than the base-case values for areas with a larger proportion of older switching equipment. In sensitivity analysis, we use costs that are twice as high: $4.20 per line and $0.002 per call.

The measurement system that Beauvais costed out will perform some functions in addition to measuring and billing local calls. For example, it will measure toll calls, and it will substitute for some traffic engineering measurements that are now done separately. The base-case cost figures take no credit for reducing the cost of these additional functions. In sensitivity analysis, we also use lower costs: $1.05 per line plus $0.0005 per call, and also zero (costless measurement). Besides being a strong lower bound on measurement costs, the zero-cost case is of some interest from a theoretical point of view—it shows what a perfect market with zero transactions costs could achieve.

PRICE RATIONING

To calculate the loss of consumer benefit due to price rationing, one needs to know a large number of own- and cross-price demand parameters, preferably at least hour by hour throughout the day. Because existing empirical estimates fall far short of that level of detail, we

5The system is capable of producing call-detail records prospectively on request for selected customers, but not retrospectively for everyone.
specified a simplified time-of-day demand system and calibrated it to be consistent with the limited information that is available.

We extended demand equations estimated in Park, Wetzel, and Mitchell (1983) to include cross-price effects and calibrated the resulting equations using several combinations of own- and cross-price effects that match the results of the GTE local measured service experiment. The patterns of response observed in the GTE experiment could have resulted from low own-price effects paired with low cross-price effects: alternatively, they could have resulted from higher values of both. To choose a "most realistic" combination of effects for use in our base-case simulation, we compared patterns of response in our model with those estimated in a major experiment with residential electricity time-of-day rates (Kohler and Mitchell, 1984). Appendix B describes the demand model in detail, reports the calibration procedure, and discusses the model's performance.

We work with a set of hourly demand curves—a residential curve and a business curve for each of 8736 hours of the year. The demand curves express the number of calls each hour as an exponential function of hourly call prices; the price during the same hour appears with a negative coefficient, and the prices during other hours appear with positive coefficients. We assume that hourly demand on any particular day is not affected by prices on other days. The equations are calibrated so that when all prices are zero, the number of calls each hour matches the number actually observed in Clinton during 1976.

The residential own-price coefficients are based directly on Park, Wetzel, and Mitchell's (1983) estimates. Business own-price coefficients are proportional to the residential coefficients, scaled down to reflect the observed relative effect of the experimental prices on business and residential calls (Jensik, 1979). Table 2 shows the own-price coefficients for the peak, shoulder, and offpeak periods that were used in the Illinois experiment, and for each of three assumed levels of cross-elasticity. The coefficients have been adjusted to compensate for the increase in the general price level between 1978 and 1984. The peak-period residential own-price coefficient $b_{pp} = -0.020$, for example, indicates that a one cent increase in peak-period price would result in a 2 percent decrease in peak-period calls.

We characterize the cross-price effects in terms of what we call "shift-out" factors. A price that applies during one hour only will reduce the number of calls during that hour by some amount, say 100. A shift-out factor of .5 means that 50 of those 100 calls are not

---

61978 prices were used in the original estimates, and the adjusted coefficients apply to prices expressed in 1984 cents.
Table 2
OWN-PRICE COEFFICIENTS IN CALIBRATED DEMAND MODEL

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>None\textsuperscript{a}</th>
<th>Medium\textsuperscript{b}</th>
<th>High\textsuperscript{c}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_{pp}$</td>
<td>-0.020</td>
<td>-0.035</td>
<td>-0.073</td>
</tr>
<tr>
<td>$b_{ss}$</td>
<td>-0.025</td>
<td>-0.042</td>
<td>-0.091</td>
</tr>
<tr>
<td>$b_{oo}$</td>
<td>-0.041</td>
<td>-0.072</td>
<td>-0.150</td>
</tr>
<tr>
<td>Business</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_{pp}$</td>
<td>-0.008</td>
<td>-0.015</td>
<td>-0.044</td>
</tr>
<tr>
<td>$b_{ss}$</td>
<td>-0.010</td>
<td>-0.020</td>
<td>-0.058</td>
</tr>
<tr>
<td>$b_{oo}$</td>
<td>-0.017</td>
<td>-0.032</td>
<td>-0.094</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Shift-out factor = 0.0.
\textsuperscript{b}Shift-out factor = 0.5.
\textsuperscript{c}Shift-out factor = 1.0.

discouraged entirely but are instead shifted to other (unpriced) hours of the same day. We calibrated our base-case demand equations using a shift-out factor of .5. This value is roughly consistent with the degree of cross-elasticity observed in an electricity pricing experiment (Kohler and Mitchell, 1984).

We do not account for complementarity between the demand for calls and the demand for access (subscription) to the telephone network. We assume that the demand for access is sufficiently less elastic than is the demand for use so that the quantitative effect of ignoring the access market is minimal (Taylor, 1980; Perl, 1984).

Sensitivity Analysis

We are confident that our base-case assumptions reasonably represent actual demand conditions. Nevertheless, we do four sensitivity analyses that vary demand assumptions.

In a high elasticity case, we double all price coefficients, including both own- and cross-price coefficients. In a low elasticity case, we halve all of the coefficients.

For a high cross-elasticity case, we recalibrate the set of demand equations using a shift-out factor of 1.0. This case serves as a probable upper bound on the possible degree of cross-elasticity; with cross-elasticity this high, local calls at different hours are perfect substitutes. A price during a single hour would not reduce total calling at all but
would simply shift some calls to other hours of the day. For a no cross-elasticity case, we recalibrate the equations using a shift-out factor of 0.0—a true lower bound on the degree of cross-elasticity.

Calculating Loss in Consumer Benefit Due to Price Rationing

Introduction of positive usage charges where flat monthly rates have been in effect will reduce the number of calls that are made and hence the total value that consumers get from making calls. The loss in consumer benefit due to price rationing is the difference between (1) consumer benefit when monthly flat rates are in effect (that is, usage prices are zero during all hours) and (2) consumer benefit when positive usage prices are in effect during some or all hours.

We calculate the loss as the sum of the areas under demand curves for each hour, measured between quantities consumed at positive prices and quantities consumed at zero prices. The demand curves, which show the relationships between hourly calls and own prices when all (own- and cross-) prices are changed in proportion, are the kind that economists call mutatis mutandis demand curves. See Appendix C for details.

QUANTITY RATIONING

Value Based on Demand Curves

The conventional approach to calculating losses due to quantity rationing is also based on demand curves. We could assume, for example, that unserved calls are randomly rationed, and evaluate the loss due to quantity rationing in each hour based on the average value of offered calls during that hour. With our assumed demand curves, the average peak-period call under flat rates is worth about 49 cents for residences and $1.21 for businesses. If positive prices are in effect, the values will be higher.

There are two problems with the conventional approach. First, the calculated value of calls lost due to quantity rationing is heavily dependent on the shape of assumed demand curves, especially their shape in high price regions where we have no empirical evidence at all. (In contrast, benefit losses due to price rationing depend on the demand curves only in the neighborhood of observed prices.) Second, the values of lost calls based on this approach imply optimal capacities.

---

7The average value of a call, calculated as the area under the demand curve divided by the number of calls, is equal to the reciprocal of the own-price coefficient, in this case \( \frac{1}{b_{pp}} \).
that are substantially lower than current telephone company practice dictates. Consequently, we explore an alternative method that places a higher value on calls lost because of quantity rationing.

**Constant Value Implied by the Engineering Capacity Rule**

If we assume that the engineering rule given above is an *optimal* way to choose capacity under flat rates, we can then deduce the implied value of rationed calls. We have already noted that, if we use Clinton 1976 data as the design load, the engineering rule chooses a capacity equal to 1533 calls per hour, a level that is exceeded during 11 hours of the year. One additional call could be served during each of those 11 hours at a capacity cost of $100 (on base-case assumptions). If the 11 additional calls were together worth more than $100, expanding capacity would be a good idea. Because the telephone company chooses (optimally, we assume) not to expand capacity to serve the additional calls, an average lost call must be worth less than $100/11 = $9.09. Also, because the company does choose to serve all calls during the tenth highest hour, a lost call must be worth more than $100/10 = $10.00. We ignore the $0.81 discontinuity and work with the former value.

**Increasing Value Implied by the Engineering Capacity Rule**

The discussion above implicitly assumes that all quantity-rationed calls are worth the same amount regardless of how overloaded the system is. But it may be more reasonable to assume that the value of serving an additional call is an increasing function of the amount of overload. If only one call goes unserved during a particular hour, the loss may be only equal to the value of that call calculated from the area under the demand curve—49 cents on average for residential calls. But as congestion increases, so do irritation and complaints; an additional unserved call may add much more than 49 cents to the social loss due to quantity rationing.

Thus, we are led to assume that the marginal value of a lost call is proportional to system overload each hour:

\[ V_i = b X_i \]

where \( V_i \) is the marginal value of a call lost during the \( i \)th quantity-rationed hour and \( X_i \) is the excess of offered load over capacity during the \( i \)th rationed hour. We can then calculate the slope coefficient \( b \) implied by telephone company practice using Clinton data as
\[ b = \$100 / \sum_{i=1}^{11} x_i = 0.0601 \]

The corresponding total quantity rationing loss during hour \( i \) may then be calculated as

\[ \int_0^x b dx = b x^2 / 2 \]

**Effects of Redialing**

We have been writing as though calls attempted in excess of hourly capacity are permanently lost. In fact, some of them are, but others are continuously redialed until they finally get through, and some are successfully retried later in the day. Redialing can affect the valuation of excess calls in two offsetting ways. On one hand, redialing may cause increasing blockage and delays as system overload increases, reinforcing our assumption that marginal quantity rationing loss is an increasing function of excess demand. On the other hand, redialing will certainly be more persistent for the more valuable calls, thus limiting the losses due to insufficient capacity.

Redialing affects the interpretation of our calculations, but it does not change the calculations themselves. When redialing is taken into account, “lost call” means a call that cannot be served anytime during the hour when it is first placed, whether it is lost entirely or merely delayed until a later hour.

**Sensitivity Analysis**

For our base-case assumption, we use a quantity rationing value that increases as the degree of system overload increases, and we calculate that value based on the assumption that the engineering capacity rule is an optimal way of choosing capacity.

In sensitivity analyses, we explore the implications of adopting the other two ways of valuing lost calls: a constant value based on the assumed optimality of the engineering capacity rule, and the average value of offered calls calculated from the area under our assumed demand curves.
USING THE MODEL

We wrote a computer program (Appendix D) to do the welfare calculations for the model described in this section. The model can be used in either a "directed" mode or a "search" mode. In the directed mode, it will calculate the components of welfare for any specified price and capacity choices. In the search mode, it does a grid search over a specified set of tariffs to locate the one that produces the maximum welfare. We used the model in both modes to obtain the results in the next section.
V. RESULTS

We used the model described above to evaluate the welfare effects of a variety of tariffs for local telephone calls, including both price and capacity choices. We first describe in detail the results under our base-case assumptions, then summarize how the results differed when the assumptions were varied in sensitivity analyses.

BASE-CASE ASSUMPTIONS FOR TECHNOLOGY IN COMMON USE TODAY

Table 3 summarizes our results for the base case, conditions that roughly represent costs for technologies in common use today. Welfare effects are measured relative to a hypothetical "benchmark" tariff with neither price nor quantity rationing.

Tariff N: A No-Rationing Benchmark

The hypothetical no-rationing tariff is shown on the first line of the table as Tariff N (a mnemonic for “no-rationing”). For this tariff, we assume (1) flat monthly rates are in effect (no additional charge for telephone calls) and (2) capacity is chosen to handle the absolute peak hourly number of calls (1968) during the year. The first assumption is representative of the way most local U.S. telephone service is now sold; the second is a convenient standard for comparison, but it is unrealistic, in that capacity is ordinarily designed to serve less than the absolute peak number of calls.

As a consequence of the two assumptions, there is neither price rationing nor quantity rationing in the benchmark tariff. Thus, the benchmark losses shown in Table 3 in columns (3)—price rationing—and (4)—quantity rationing—are zero. Because we measure capacity savings relative to the annual cost of benchmark capacity, the benchmark figure for capacity savings in column (1) is zero as well. Also, there is no need to keep track of the number of calls when they are priced at zero, so there are no measurement costs in column (2). The four zeros add up to a zero welfare benchmark against which the other tariffs can be measured.

All the other tariffs reported in Table 3 result in lower capacity and thus save on capacity cost relative to the benchmark tariff. As there is always a partially offsetting rationing loss, and for tariffs with positive
### Table 3
WELFARE EFFECTS OF LOCAL MEASURED TELEPHONE SERVICE IN A SMALL EXCHANGE:
BASE-CASE ASSUMPTIONS FOR TECHNOLOGY IN COMMON USE TODAY

<table>
<thead>
<tr>
<th>Tariff</th>
<th>Weekday Periods</th>
<th>Prices</th>
<th>Capacity Cost (calls per hour)</th>
<th>Capacity Cost Savings (1)</th>
<th>Measurement Costs (2)</th>
<th>Price Rationing Loss (3)</th>
<th>Quantity Rationing Loss (4)</th>
<th>Welfare Gain (5) = (1)−(2)−(3)−(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N:</td>
<td>00-24</td>
<td>.00</td>
<td>peak load</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>P:</td>
<td>various</td>
<td>.338</td>
<td>optimal</td>
<td>109.5</td>
<td>9.1</td>
<td>10.5</td>
<td>0.0</td>
<td>89.9</td>
</tr>
<tr>
<td></td>
<td>max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3:</td>
<td>11-16</td>
<td>.10</td>
<td>peak load</td>
<td>22.2</td>
<td>8.9</td>
<td>11.9</td>
<td>0.0</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>10-11,16-17</td>
<td>.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17-10</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q:</td>
<td>00-24</td>
<td>.00</td>
<td>engineered</td>
<td>43.5</td>
<td>0.0</td>
<td>0.0</td>
<td>12.4</td>
<td>31.1</td>
</tr>
<tr>
<td>Q1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1:</td>
<td>00-24</td>
<td>.04</td>
<td>engineered</td>
<td>43.5</td>
<td>0.0</td>
<td>0.0</td>
<td>12.4</td>
<td>31.1</td>
</tr>
<tr>
<td>M1:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M2:</td>
<td>10-17</td>
<td>.07</td>
<td>engineered</td>
<td>52.6</td>
<td>11.2</td>
<td>5.8</td>
<td>11.5</td>
<td>24.2</td>
</tr>
<tr>
<td>M2:</td>
<td>17-10</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3:</td>
<td>10-17</td>
<td>.09</td>
<td>engineered</td>
<td>72.2</td>
<td>9.6</td>
<td>12.1</td>
<td>10.4</td>
<td>40.2</td>
</tr>
<tr>
<td>M3:</td>
<td>08-10, 17-18</td>
<td>.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3:</td>
<td>18-06</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTES:** Columns (1) through (6) are $1000 per year. Base-case assumptions are as follows.

- **Incremental capacity cost:** $100 per year per call.
- **Measurement cost:** $2.10 per year per line + $0.0010 per call.
- **Elasticity:** Based on Park, Wietz and Mitchell (1983).
- **Cross-elasticity:** Shift-out factor = 0.5.
- **Customer mix:** Actual Clinton residential and business.
- **Quantity racioning value:** Increasing function of the number of lost calls; based on assumed optimality of the engineering capacity rule.
prices measurement costs as well, the trick is to find price and capacity choices that maximize the difference between the capacity savings and the losses. In the present case, optimal first-best hour-by-hour prices maximize welfare.¹

**Tariff P: First-Best Hour-by-Hour Prices**

The situation with optimal first-best hour-by-hour prices is shown as Tariff P (for "price-rationing"). The price during the very busiest hour of the year is 34 cents per call, which reduces offered load during that hour from 1968 calls to 873 calls. Graduated lesser prices during 3296 other busy hours reduce offered load to the same level. Capacity need only be constructed to serve 873 calls; capacity cost savings are a substantial $109,500 per year. Price rationing occurs only during hours when it is necessary to level demand to the optimal capacity level; the price rationing loss due to first-best prices is $10,500. No quantity rationing is required, so the quantity rationing loss is zero. Measuring and billing for calls made during hours with positive prices costs $9,100. Subtracting the losses from the capacity savings yields a net welfare gain over the benchmark tariff of $89,900, or about $25 per line per year.

Optimal first-best prices require perfect foreknowledge of hour-by-hour demand throughout the year; and even if such knowledge were available, fully optimal prices are far too complex to be a satisfactory guide to customer calling decisions. Too many prices in a feasible tariff will confuse the customer. Conventionally, a maximum of three prices—peak, shoulder, and offpeak—have been used in time-of-day tariffs for toll calls (and similarly for electricity), and we accept that limit here. The limit presumably strikes a rough balance between incremental efficiency gains and incremental transactions costs of more complex tariffs.

In the remaining tariffs, we shall see how closely we can approach the welfare standard set by optimal first-best prices when we are limited to feasible tariffs.

**Tariff P3: Feasible Pure Price Rationing**

Tariff P3 finds the best three prices and periods when the only rationing allowed is price rationing, so that capacity must be con-

---

¹Higher measurement costs, lower price elasticities, or lower capacity costs could conceivably make even first-best prices less efficient than pure quantity rationing in some other cases—not, however, in any of our sensitivity analysis cases.
structured to serve all of the calls offered at the specified prices. Tariff P3 is optimal subject to the following constraints:

- A single positive price applies to all peak hours;
- The shoulder and offpeak prices may be positive or zero;
- Only integer prices (cents per call) are allowed;\(^2\)
- The peak period consists of the same contiguous hours of the day on each weekday of the year;\(^3\)
- Contiguous shoulder periods are allowed on either side of the peak period;
- Capacity is chosen to serve the maximum number of calls that are offered when the prices are in place.

The best feasible pure price rationing tariff charges 10 cents for calls during an 11 a.m. to 4 p.m. peak period, 6 cents for calls during one-hour shoulder periods on either side of the peak, and zero during other (offpeak) hours.\(^4\) These prices reduce the maximum number of calls during any hour of the year to 1546; because capacity must be constructed to serve all of these calls, capacity cost savings are much lower than savings for the first-best Tariff P. Consequently, feasible pure price rationing produces a welfare gain over the benchmark non-rationing tariff of only $21,400.

**Tariff Q: Flat Rates (Quantity Rationing Only)**

The pure quantity-rationing Tariff Q (for “quantity rationing”) approximates the situation in most local exchanges in the United States today. Flat monthly rates are in effect, with no extra charge for calls, and capacity is engineered so that 1 percent of calls attempted

---

\(^2\)We adopt this constraint only to keep computation within manageable bounds. The welfare improvements to be had from fine-tuning fractional prices are minor; the welfare function seems to be rather flat in the vicinity of the optimum.

\(^3\)Tariffs that also charge peak prices on Saturdays (or Saturdays and Sundays) are less efficient for our telephone calling pattern than are tariffs with peak prices on weekdays only.

\(^4\)This is Craven’s (1985) case, because of the assumption that capacity equals the maximum number of offered calls. Our zero price during offpeak hours apparently conflicts with his theorem that the optimal offpeak price is always greater than zero. The apparent conflict arises because we do not allow fractional prices or price periods that are based on less than full hours. If we did, we could presumably find an offpeak price slightly greater than zero and a corresponding slight increase in the length of the offpeak period that would improve on Tariff P3.
during the 20 busy-season busy hours cannot be completed. This results in capacity cost savings of $43,500. In our data, there are 11 hours during the year with offered loads in excess of the Tariff Q capacity of 1533 calls. During these hours, a total of 1664 potential calls are not served, resulting in a quantity rationing loss of $12,400. This loss is a partial offset to the capacity cost savings, so the net welfare gain over the benchmark tariff is $31,100.

If one is restricted to feasible pure price rationing or (feasible) pure quantity rationing in our base-case situation, pure quantity rationing wins. However, combinations of the two may do better than either alone. The remaining "mixed" tariffs, M1 through M3, explore the use of price rationing in conjunction with quantity rationing to improve on the pure strategies.

**Tariff M1: Optimal Uniform Price**

Tariff M1 is a feasible tariff that combines price and quantity rationing. (The M is for "mixed price and quantity rationing.") For this tariff, we search out the optimal single price per call subject to the following constraints:

- A single positive price applies to all hours;
- Only integer numbers of cents per call are considered;
- Capacity is chosen according to the engineering rule described above.

Under these restrictions, the optimal price per call is 4 cents. This price reduces offered load during the busy-season busy hours (as well as during all other times of the year), hence reduces engineered capacity, yielding capacity savings over flat rate Tariff Q. More than offsetting this gain are price rationing losses and measurement costs that are not present in Tariff Q. On balance, the optimal single price improves welfare over the benchmark Tariff N by $24,200.

In our base-case situation, then, a zero price is better than a single positive price, because measurement costs more than offset what would otherwise be a modest welfare gain due to the positive price.

**Tariff M2: Optimal Two Prices and Periods**

A price that varies by time of day can do a better job of discriminating between hours when rationing is desirable and hours when excess capacity is available. Two or three prices are feasible if they apply during regular repeating hours of the day. In Tariff M2, we search out the optimal two-price tariff. Both peak and offpeak prices are allowed
to vary, and so are the hours during which they apply, subject to these constraints:

- The peak price is positive;
- The pricing periods are contiguous hours that repeat from day to day;
- Prices are integer cents;
- Capacity is chosen by the engineering rule.

The optimal two-period tariff sets a peak price of 7 cents per call and an offpeak price of zero. The peak period runs from 10 a.m. to 5 p.m. weekdays; nights, evenings, and all hours on weekends are off peak. Capacity cost savings increase substantially over the best single-price tariff, while the offsetting losses change by smaller amounts. The welfare gain relative to the benchmark Tariff N is $38,900, somewhat better than that achieved by flat rates, Tariff Q.

**Tariff M3: Optimal Three Prices and Periods**

A three-price tariff is also feasible and should be able to improve welfare further by better matching fluctuations in load. In Tariff M3, we find the optimal three-price tariff subject to:

- Positive peak price;
- Contiguous peak period;
- Contiguous shoulder (intermediate-price) periods allowed on either side of the peak;
- Integer prices;
- Engineered capacity.

In our base-case situation, the optimal three-price tariff improves slightly on the optimal two-price tariff. (In some sensitivity analysis cases, summarized below, three prices are no better than two.) One might have hoped that time-of-day tariffs with a small number of prices would quickly approach the gains obtainable with optimal hour-by-hour prices (Tariff P), but that seems not to be true here. Computational complexity mounts rapidly as the number of price periods increases, so we did not attempt to find optimal four- or five-period prices. The small (or zero) gain in going from two to three periods leads us to expect that any gains from going to four or five periods would be minimal.
SENSITIVITY ANALYSES

We did similar analyses under alternative sensitivity analysis assumptions. Most of the alternative assumptions are described in Sec. IV above; all are summarized in Table 4.

Welfare measures for the various cases (relative to the benchmark no-rationing tariff) are shown in Table 5 for the same types of tariffs that we explored for the base case in Table 3. Full details, together

Table 4

<table>
<thead>
<tr>
<th>Base-case Assumptions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incremental capacity cost</td>
<td>$100 per year per call</td>
</tr>
<tr>
<td>Measurement cost</td>
<td>$2.10 per year per line + $0.0010 per call</td>
</tr>
<tr>
<td>Elasticity</td>
<td>Based on Park, Wetzel and Mitchell (PWM, 1983)</td>
</tr>
<tr>
<td>Cross-elasticity</td>
<td>Shift-out factor = 0.5</td>
</tr>
<tr>
<td>Customer mix</td>
<td>Actual Clinton residential and business</td>
</tr>
<tr>
<td>Quantity rationing value</td>
<td>Increasing function of the number of lost calls;</td>
</tr>
<tr>
<td></td>
<td>based on assumed optimality of the engineering capacity rule</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sensitivity Analysis Cases</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incremental capacity cost</td>
<td></td>
</tr>
<tr>
<td>Higher</td>
<td>$200 per year per call</td>
</tr>
<tr>
<td>Lower</td>
<td>$50 per year per call</td>
</tr>
<tr>
<td>Much lower</td>
<td>$20 per year per call</td>
</tr>
<tr>
<td>Measurement cost</td>
<td></td>
</tr>
<tr>
<td>Higher</td>
<td>$4.20 per line + $0.0020 per call</td>
</tr>
<tr>
<td>Lower</td>
<td>$1.05 per line + $0.0005 per call</td>
</tr>
<tr>
<td>Zero</td>
<td>$0.00 per line + $0.0000 per call</td>
</tr>
<tr>
<td>Elasticity</td>
<td></td>
</tr>
<tr>
<td>Higher</td>
<td>Twice PWM estimates</td>
</tr>
<tr>
<td>Lower</td>
<td>Half PWM estimates</td>
</tr>
<tr>
<td>Higher cross-elasticity</td>
<td>1.0 shift-out factor</td>
</tr>
<tr>
<td>No cross-elasticity</td>
<td>0.0 shift-out factor</td>
</tr>
<tr>
<td>Customer mix</td>
<td></td>
</tr>
<tr>
<td>Residential exchange</td>
<td>90 percent of total calls are by residences</td>
</tr>
<tr>
<td>Business exchange</td>
<td>90 percent of total calls are by businesses</td>
</tr>
<tr>
<td>Quantity rationing value</td>
<td></td>
</tr>
<tr>
<td>Constant rationing value</td>
<td>Value of an incremental call lost to quantity rationing is</td>
</tr>
<tr>
<td></td>
<td>a constant based on assumed optimality of the engineering capacity rule</td>
</tr>
<tr>
<td>Demand rationing value</td>
<td>Value of an incremental call lost to quantity rationing is</td>
</tr>
<tr>
<td></td>
<td>based on area under demand curves.</td>
</tr>
</tbody>
</table>
with results for a few additional tariffs, are in Appendix Tables E.1 to E.15.

The major results for the base case are generally true in the sensitivity analyses as well:

1. Feasible pure price rationing (P3) is almost always less efficient than (feasible) pure quantity rationing (Q).
2. Unless capacity costs are very high or measurement costs very low, the best single price (M1) is less efficient than a flat monthly rate with no usage charge (Q).
3. Unless capacity costs are low or measurement costs are quite high, the best two-period prices (M2) provide modest efficiency gains over flat rates.
4. Without exception, the best three-period prices (M3) improve on two-period prices only slightly, or not at all.
5. The best feasible time-of-day prices (M3) fall far short of the efficiency gains that are theoretically attainable with first-best hour-by-hour prices (P).

Table 5
SENSITIVITY ANALYSIS WELFARE RESULTS
($1000)

<table>
<thead>
<tr>
<th>Case</th>
<th>Hourly Prices</th>
<th>Pure Price</th>
<th>Flat Rate</th>
<th>One Price</th>
<th>Two Prices</th>
<th>Three Prices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>P3</td>
<td>Q</td>
<td>M1</td>
<td>M2</td>
<td>M3</td>
</tr>
<tr>
<td>Base-case assumptions</td>
<td>89.9</td>
<td>25.3</td>
<td>31.1</td>
<td>24.2</td>
<td>38.9</td>
<td>40.2</td>
</tr>
<tr>
<td>Higher capacity cost</td>
<td>204.1</td>
<td>73.8</td>
<td>62.2</td>
<td>68.6</td>
<td>100.4</td>
<td>108.1</td>
</tr>
<tr>
<td>Lower capacity cost</td>
<td>37.2</td>
<td>3.8</td>
<td>15.5</td>
<td>5.3</td>
<td>12.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Much lower capacity cost</td>
<td>8.2</td>
<td>-4.3</td>
<td>6.2</td>
<td>-5.1</td>
<td>-1.1</td>
<td>-1.1</td>
</tr>
<tr>
<td>Higher measurement cost</td>
<td>80.8</td>
<td>16.4</td>
<td>31.1</td>
<td>12.9</td>
<td>30.0</td>
<td>30.6</td>
</tr>
<tr>
<td>Lower measurement cost</td>
<td>94.4</td>
<td>29.8</td>
<td>31.1</td>
<td>29.8</td>
<td>43.4</td>
<td>44.9</td>
</tr>
<tr>
<td>Zero measurement cost</td>
<td>99.0</td>
<td>34.7</td>
<td>31.1</td>
<td>35.4</td>
<td>48.3</td>
<td>49.7</td>
</tr>
<tr>
<td>Lower capacity cost and</td>
<td>41.4</td>
<td>8.3</td>
<td>15.5</td>
<td>10.9</td>
<td>16.5</td>
<td>16.5</td>
</tr>
<tr>
<td>lower measurement cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Much lower capacity cost</td>
<td></td>
<td>18.3</td>
<td>-0.2</td>
<td>4.5</td>
<td>0.6</td>
<td>2.5</td>
</tr>
<tr>
<td>and lower measurement cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher elasticity</td>
<td>97.3</td>
<td>32.2</td>
<td>31.1</td>
<td>28.7</td>
<td>44.5</td>
<td>48.3</td>
</tr>
<tr>
<td>Lower elasticity</td>
<td>82.9</td>
<td>16.6</td>
<td>31.1</td>
<td>21.9</td>
<td>32.9</td>
<td>33.0</td>
</tr>
<tr>
<td>Higher cross-elasticity</td>
<td>95.4</td>
<td>27.5</td>
<td>31.1</td>
<td>21.6</td>
<td>39.6</td>
<td>45.4</td>
</tr>
<tr>
<td>No cross-elasticity</td>
<td>85.6</td>
<td>21.6</td>
<td>31.1</td>
<td>25.4</td>
<td>37.1</td>
<td>37.1</td>
</tr>
<tr>
<td>Residential exchange</td>
<td>89.2</td>
<td>24.9</td>
<td>29.2</td>
<td>22.1</td>
<td>35.0</td>
<td>38.3</td>
</tr>
<tr>
<td>Business exchange</td>
<td>116.0</td>
<td>40.9</td>
<td>63.3</td>
<td>58.3</td>
<td>66.3</td>
<td>67.6</td>
</tr>
<tr>
<td>Constant rationing value</td>
<td>89.9</td>
<td>25.3</td>
<td>28.4</td>
<td>21.1</td>
<td>33.4</td>
<td>35.0</td>
</tr>
<tr>
<td>Demand rationing value</td>
<td>89.9</td>
<td>25.3</td>
<td>42.8</td>
<td>35.1</td>
<td>64.8</td>
<td>66.0</td>
</tr>
</tbody>
</table>
VI. CONCLUSION

We have reached two major conclusions about efficient time-of-day pricing for services with important capacity costs. The first has general implications for the application of peak-load pricing theory; the second is specific to public policy toward pricing of local telephone service.

1. Variations in demand within pricing periods may sharply limit the efficiency gains that price rationing can achieve. In important cases, the requirement that pricing be feasible (with a small number of different prices and periods limited to regular weekday and weekend hours) makes any feasible prices by themselves less efficient than a zero price with some amount of quantity rationing. A combination of quantity rationing with prices that have been carefully designed in light of local demand and cost conditions may sometimes be the best feasible tariff. Because intra-period demand variation can substantially affect the relative efficiency of alternative tariffs, it should be accounted for in applications of peak-load pricing theory to other industries.

2. Measured-rate pricing of local telephone calls is unlikely to increase economic efficiency in most circumstances, even if the tariffs are carefully designed and the costs of measuring calls are minimal. The demand for local telephone calls varies substantially within feasible time-of-day pricing periods, and the marginal costs of additional capacity have probably fallen below our base-case values with the introduction of digital electronics and optical transmission. In these circumstances, a flat monthly rate with no extra charge for use (a tariff that rations demand during just a few of the highest use hours of the year) is apt to be at least as efficient as feasible price rationing (which rations demand during many hours when excess capacity is available). Efficiency considerations alone do not support a public policy choice of measured rates over flat rates for local telephone service.

APPLIED PEAK-LOAD PRICING THEORY

Our base-case results, summarized in Table 6, establish the importance of intra-period demand variation in one application of peak-load pricing theory.

Tariff P charges separate, graduated prices during each of 3297 high-demand hours of the year to level out peak traffic to the optimal
Table 6
SUMMARY OF BASE-CASE RESULTS FOR TECHNOLOGY IN COMMON USE TODAY

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N: Benchmark</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>0.0</td>
</tr>
<tr>
<td>P: First-best prices</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>89.9</td>
</tr>
<tr>
<td>P3: Feasible pure price rationing</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>25.3</td>
</tr>
<tr>
<td>Q: Flat rates</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>31.1</td>
</tr>
<tr>
<td>M3: Feasible price and quantity rationing</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>40.2</td>
</tr>
</tbody>
</table>

capacity of the telephone exchange. Tariff P achieves a large welfare gain because all rationing is precisely targeted: The hourly prices discourage only the lowest valued calls each hour, and the prices apply only during those hours when unrationed demand exceeds optimal capacity.

Such precise targeting is not feasible; in practice, we are limited to perhaps three regular price periods. With feasible prices and a smoothly downward sloping load-duration curve for unrationed yearly traffic (Fig. 3), intra-period demand variation is unavoidable. Then the welfare gains attainable by pricing alone (Tariff P3, which we consider solely for its theoretical interest) are sharply curtailed—only 28 percent of the gains achieved by first-best prices. Feasible prices still eliminate only the lowest valued calls each hour, but they fail to level the load at optimal capacity and they apply during many hours when excess capacity is available.

Quantity rationing (Tariff Q, the most common way of pricing local telephone service today) targets the busiest hours of the year and thus avoids rationing during hours when rationing is unnecessary. But during rationed hours, it indiscriminately affects both low and high valued calls. Pure quantity rationing does better than pure price rationing in our base case, achieving about 35 percent of the welfare gain theoretically achievable by first-best prices.

Mixed price and quantity rationing (Tariff M3, a feasible mandatory measured service tariff) does somewhat better than either price or
quantity rationing alone, yielding about 45 percent of the welfare gain from first-best prices.

The fact that demand varies substantially within any feasible price periods has two major implications in our base-case situation (Table 6), in our sensitivity analyses (Table 5), and, we suspect, in other applications of peak-load pricing theory as well:

- The welfare gains achievable by feasible prices are sharply limited, and
- Quantity rationing (or a mixture of price and quantity rationing) is more efficient than are feasible prices alone.

PUBLIC POLICY TOWARD TELEPHONE PRICING

Feasible time-of-day prices (M3) yielded a welfare gain ($40,200 – $31,100 = $9,100) over flat rate prices (Q) in our base-case situation. There are at least two reasons to believe that this modest gain is an upper bound on the efficiency gains that can be expected of local measured service telephone rates in practice, and that measured rates are in fact more likely to result in small reductions in economic efficiency.

1. The digital switching and fiber optic transmission systems that will be the predominant technologies installed over the coming decade have incremental capacity costs that are expected to be substantially lower than the base-case value. The future situation is probably better approximated by our "lower" and "much lower" capacity cost assumptions.\(^1\) In neither of these cases are the best prices able to increase welfare over flat rates.

2. Feasible prices achieve a modest gain over a flat rate in the base-case situation because time-of-day rates and periods were carefully tailored to particular local traffic and cost conditions. In practice, measured service tariffs are unlikely to be optimal in any particular locality. One reason is that considerations other than economic efficiency may legimitely influence tariff design. Also, in practice, the same tariff would have to apply over large areas, probably an entire state; then geographical variations in traffic and costs will undermine the improvement that measured service might achieve in any single locality.

We explored this phenomenon in a limited way in Appendix E by evaluating some tariffs that are not optimally tailored for local conditions. In contrast to optimal tariffs, which generally charge nothing for calls during offpeak hours, most actual tariffs charge positive

\(^1\)Appendix Tables E.3 and E.4.
(though discounted) prices during the evening and night. One such tariff is less efficient than flat rates in most of our sensitivity analysis situations.\textsuperscript{2} We also evaluated the welfare achieved by the best base-case feasible prices and periods (M3 in Table 3), when they are applied to other situations.\textsuperscript{3} We found that the best base-case tariff was usually only slightly inferior to the optimal tariff in our sensitivity analysis situations. Larger differences in traffic patterns would, of course, lead to larger efficiency differences.

Measured service pricing of local telephone calls has appeared attractive to telephone companies seeking automatic revenue increases with rising telephone use, to regulators hoping to maintain a low monthly charge for access to telephone service, and to economists who would like to see scarce resources used more efficiently. We find that, contrary to conventional wisdom, measured rate pricing of local telephone calls is likely to be somewhat less efficient than traditional flat rate pricing. If local measured service is desirable public policy, it must be justified on grounds other than economic efficiency.

\textsuperscript{2}Appendix Tables E.1 to E.17, Tariff M5.
\textsuperscript{3}Appendix Tables E.2 to E.17, Tariff M0.
VII. POSTSCRIPT

Publication of the original version of this report in June 1986 has stimulated discussion of local telephone pricing, subjected this study to critical review, and suggested the need for further research.\footnote{See Beauvais (1986), Koschat, Lehman, and Sieff (1986), and Shew (1987).} We wish to clarify several points in the report to focus further efforts in productive areas.

1. We concluded that measured service is apt to be somewhat less efficient than flat rate pricing for local telephone service. Several discussants have pointed out that our base-case analysis, along with most of the other cases in Table 5, fails to support that conclusion. Indeed, measured service yields higher welfare than flat rates in the base case. However, our policy conclusion is not derived from the base case itself, but rather from some of the more realistic sensitivity analyses.

The base case uses publicly available incremental cost estimates, not because we believe that they represent the costs of the digital switching and fiberoptic trunking facilities that are currently being installed, but simply because they are publicly available. Some unpublished studies by suppliers of telecommunications services suggest that current costs are substantially lower than the base-case costs—perhaps even lower than in our lower-cost and much-lower-cost cases, even in local calling areas with several switching offices.

We based our policy conclusion on the sensitivity analysis cases that in our judgment come closest to representing the current incremental costs of expanding capacity and measuring calls with modern technology. Additional research on telephone costs and open discussion of the research will be needed to confirm or refute that judgment.

2. Some discussants have stressed that even if flat rates can be more efficient than the best feasible measured rates, the efficiency differences are not large. We agree. In our sensitivity analysis, the flat-rate efficiency advantage over measured service never exceeds about $2 per line per year. This may be a large sum when multiplied by the number of telephone lines in the United States, but it is not large compared with the remaining uncertainty about the representativeness of our results.

In such circumstances, policy choices between measured service and flat rates will, and rightly should, be based in large measure on how well they serve public policy goals other than economic efficiency.
3. If, as a society, we do opt for measured rates, tariffs should be carefully designed to avoid large, unnecessary efficiency losses. In our examples, the best measured-service rates charge nothing for calling during a large number of off-peak hours. Most actual (existing and proposed) measured-service tariffs charge for calls at all times of the day and night. Results presented in Appendix E show that charging for calls when substantial excess capacity is available can carry a large efficiency penalty.

4. We did not analyze "feasible" tariffs any more complex than mandatory rate structures with three distinct prices for calls. Several discussants have suggested that more complex tariffs are, in fact, feasible and might be substantially more efficient than those we analyzed. We agree that a limit of three prices is not hard and fast, but we are somewhat skeptical about the prospects for large gains from more complex tariffs. The reason for our skepticism is that the additional efficiency gained by allowing three rather than two prices is never very large in our examples.

Nevertheless, it is conceivable that a careful analysis of more complex tariffs might demonstrate the possibility of larger gains. Candidates for further study include

- a larger number of different prices during the day,
- seasonally varying prices,
- different tariffs for different classes of customers,
- separate charges for calls and for call duration,
- optional measured rates,
- tariffs with some free calls and an upper limit on the total monthly bill.

5. Our analysis neglects welfare effects in other markets than that for local telephone calls. Some discussants have suggested that broadening the analysis to include access (monthly subscriptions to telephone service) and long distance calls would change our conclusions. We acknowledge the possibility but believe that the arguments need substantial development before they will be very convincing.

It is widely believed that prices of long distance calls substantially exceed marginal costs and that monthly charges for access to the telephone network are well below incremental costs. If so, repricing those services could increase welfare. However, those price changes do not require that local telephone service be priced at measured rates; the efficiency of local measured service itself does not turn on the gains that could be achieved in the long distance market.
If measured service is introduced with no change in other prices, the number of subscribers will decline; if the fixed monthly charge is reduced to offset the revenue from per-call charges, the number of subscribers may increase. The welfare effects of a change in subscribers depends on the relationship of the access price and the marginal cost of access, net of any external benefits existing subscribers receive from the marginal subscribers. If the access price is currently below social marginal cost, the reduction in access price that will accompany the introduction of measured service will cause a welfare loss in the access market.

6. Some discussants stress that our traffic data are not directly representative of urban and metropolitan conditions. We attempted to cover a wide spectrum of possible traffic patterns by varying the mix of residence and business traffic in sensitivity analyses. But clearly, real data from other localities should be analyzed to determine to what extent our findings hold up in other situations.
BIBLIOGRAPHY


